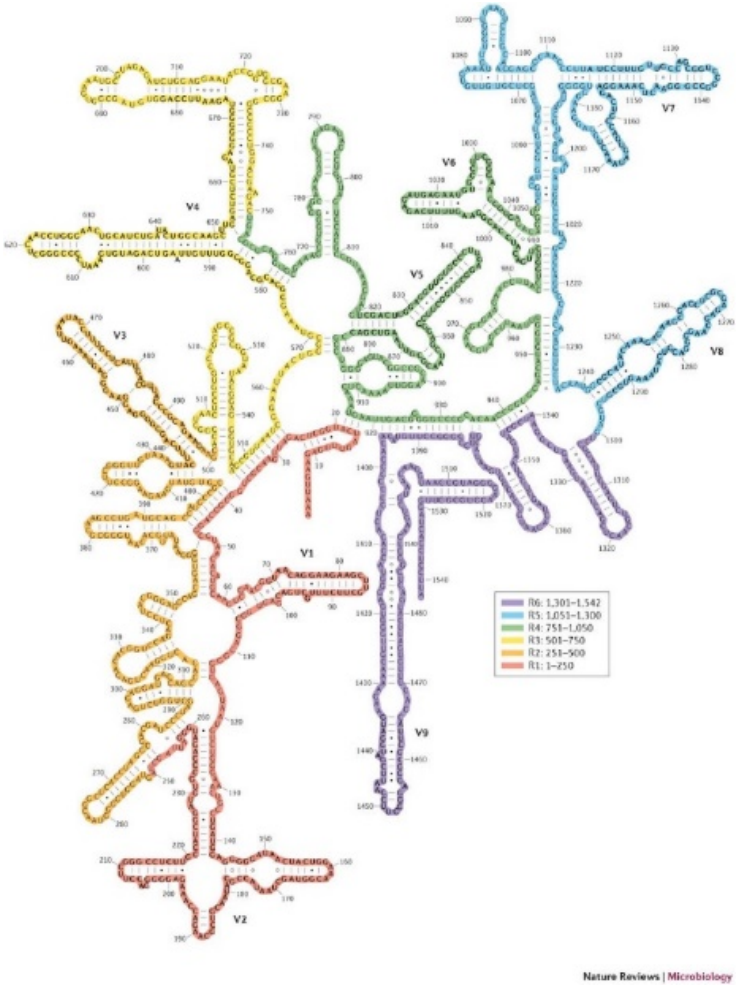
Day 4:

Metataxonomic analysis - *The Amplicon Session*

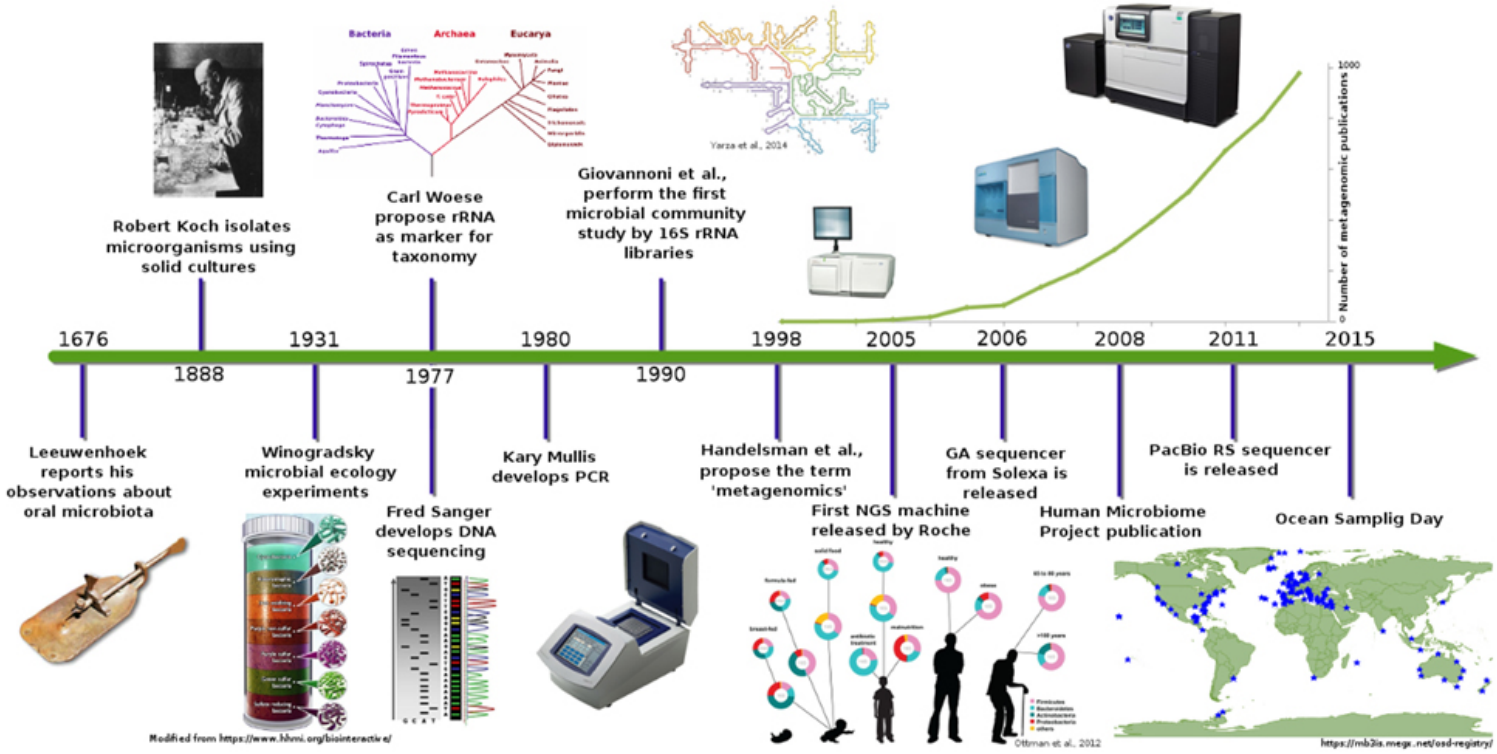


Tomáš Větrovský

Laboratory of Environmental Microbiology

Targeted gene sequencing

- focusing on specific genes
- can target a large diversity of organisms
- amplification of the needed gene area via PCR followed by High-Throughput Sequencing (HTC)

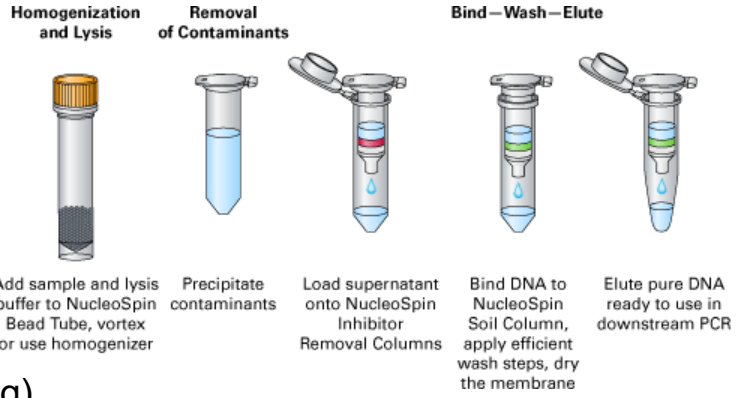


…use of HTS reveals huge diversity of un-culturable or previously unknown microorganisms (only around 1-2% of bacteria can be cultured in the laboratory)

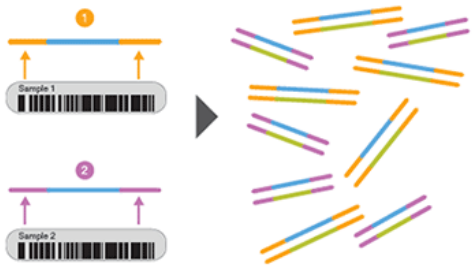
Escobar-Zepeda (2015)The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. Front. Genet. https://doi.org/10.3389/fgene.2015.00348

Illumina amplicon sequence library preparation

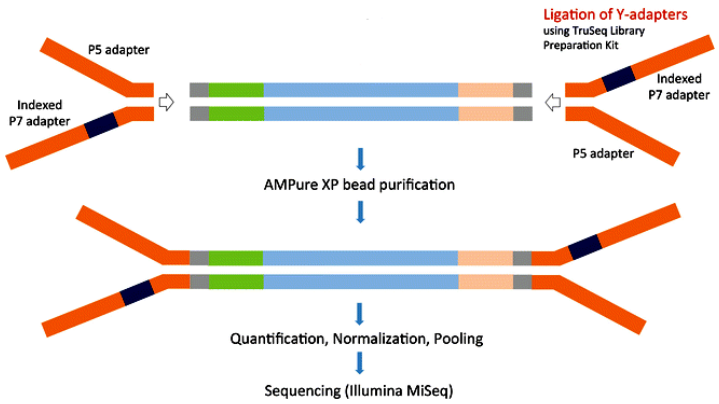
1. DNA isolation



2. PCR with barcoded primers (multiplexing)



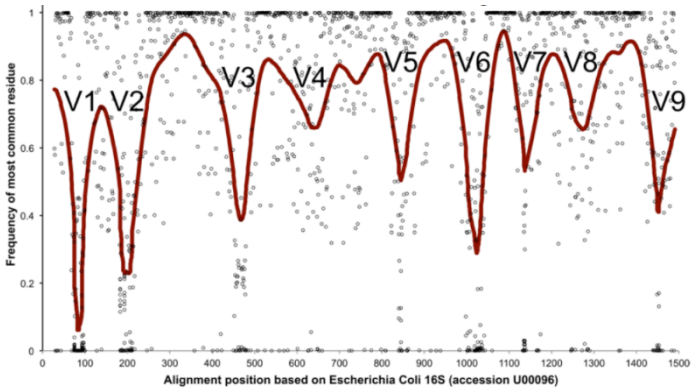
3. Ligation of sequencing adapters – to attach short oligonucleotides (60bp) to your DNA fragments, these oligonucleotides are used to attach to the sequencing flow cell and they are also used as barcode of library



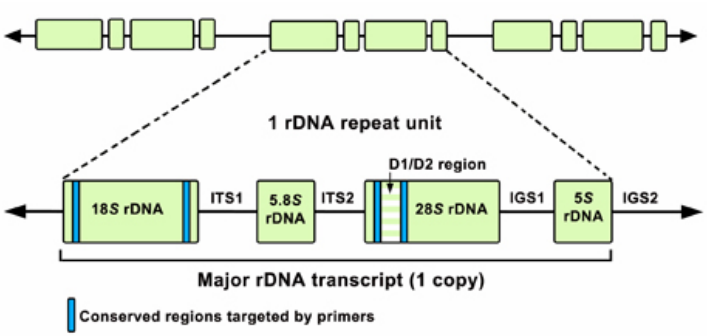
4. Quantification of the library by qPCR – to quantify of the exact amount of ligated fragments

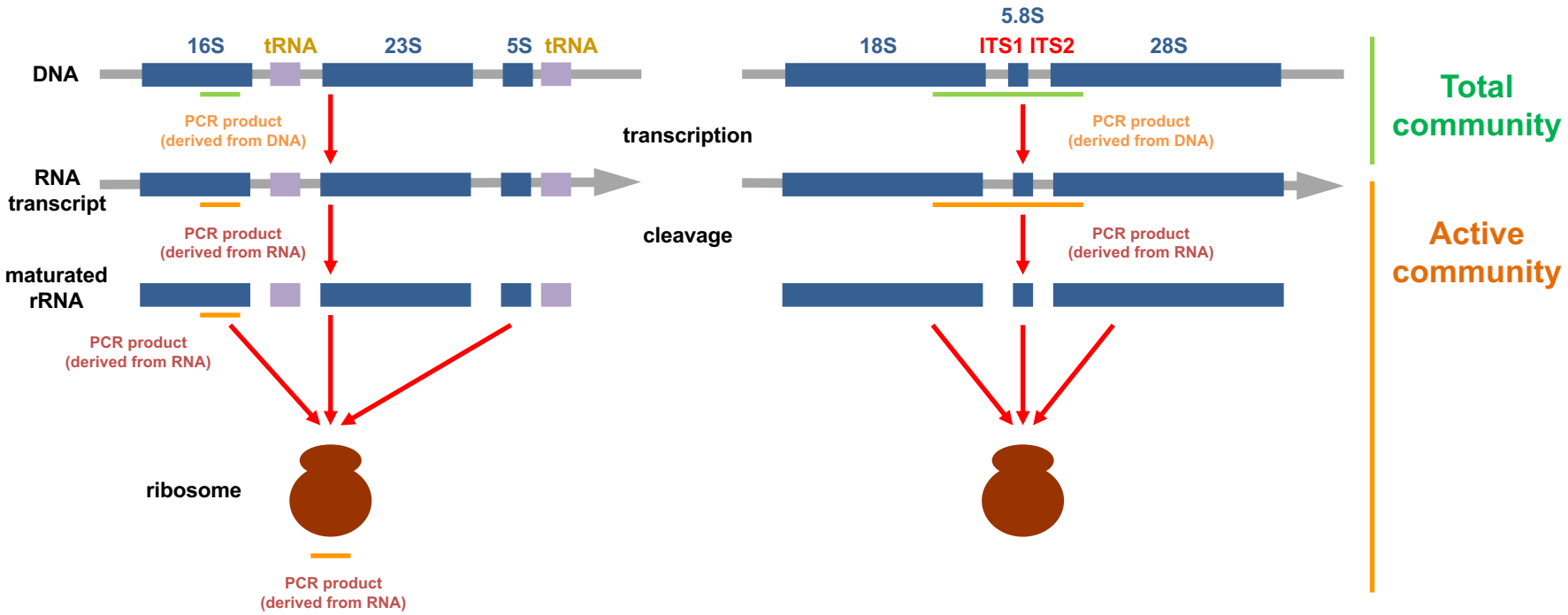
Most used marker genes

Bacterial 16S ribosomal RNA gene



Fungal internal transcribed spacer (ITS)





1-15 copies of rDNA per genome

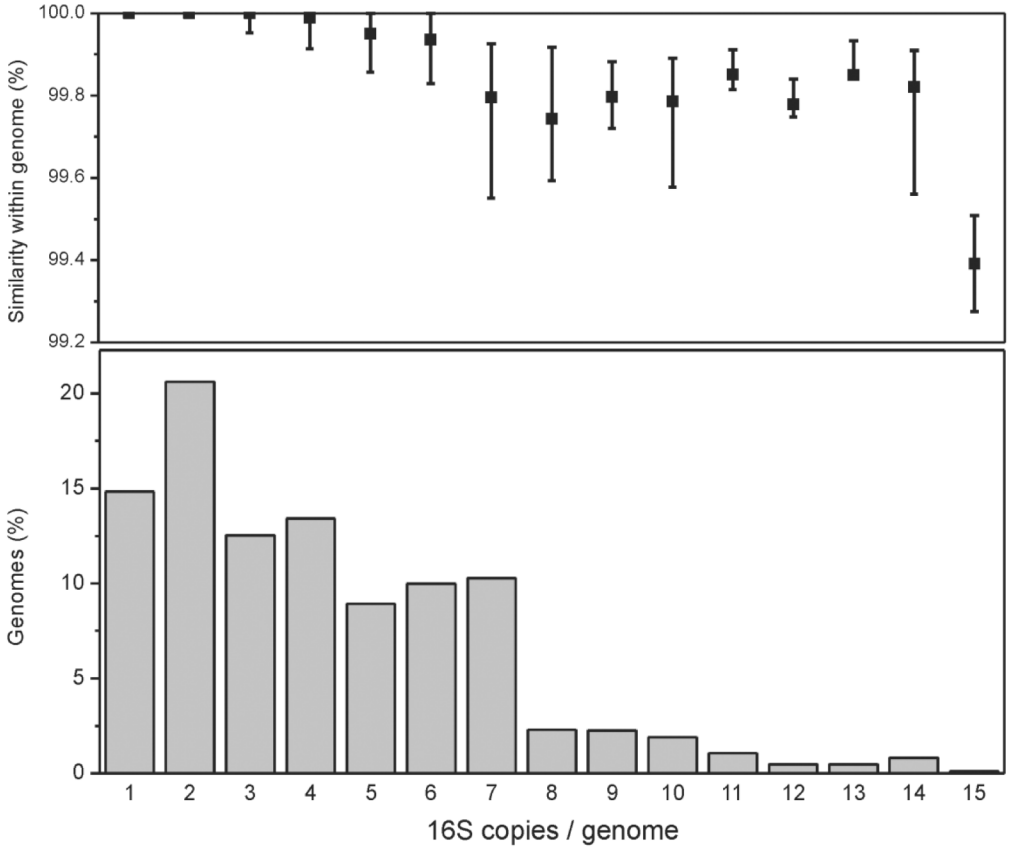
X0–X00? copies of ITS per genome

16S rDNA gene vs. alternative (low-copy) markers

Pros: highly populated reference databases

Cons: muticopy nature of bacterial 16S rDNA gene

- possibility of high intragenomic variability - diversity over estimation (number of OTUs)
- relative abundance estimation is skew -> normalisation by 16S copy number of closest taxon



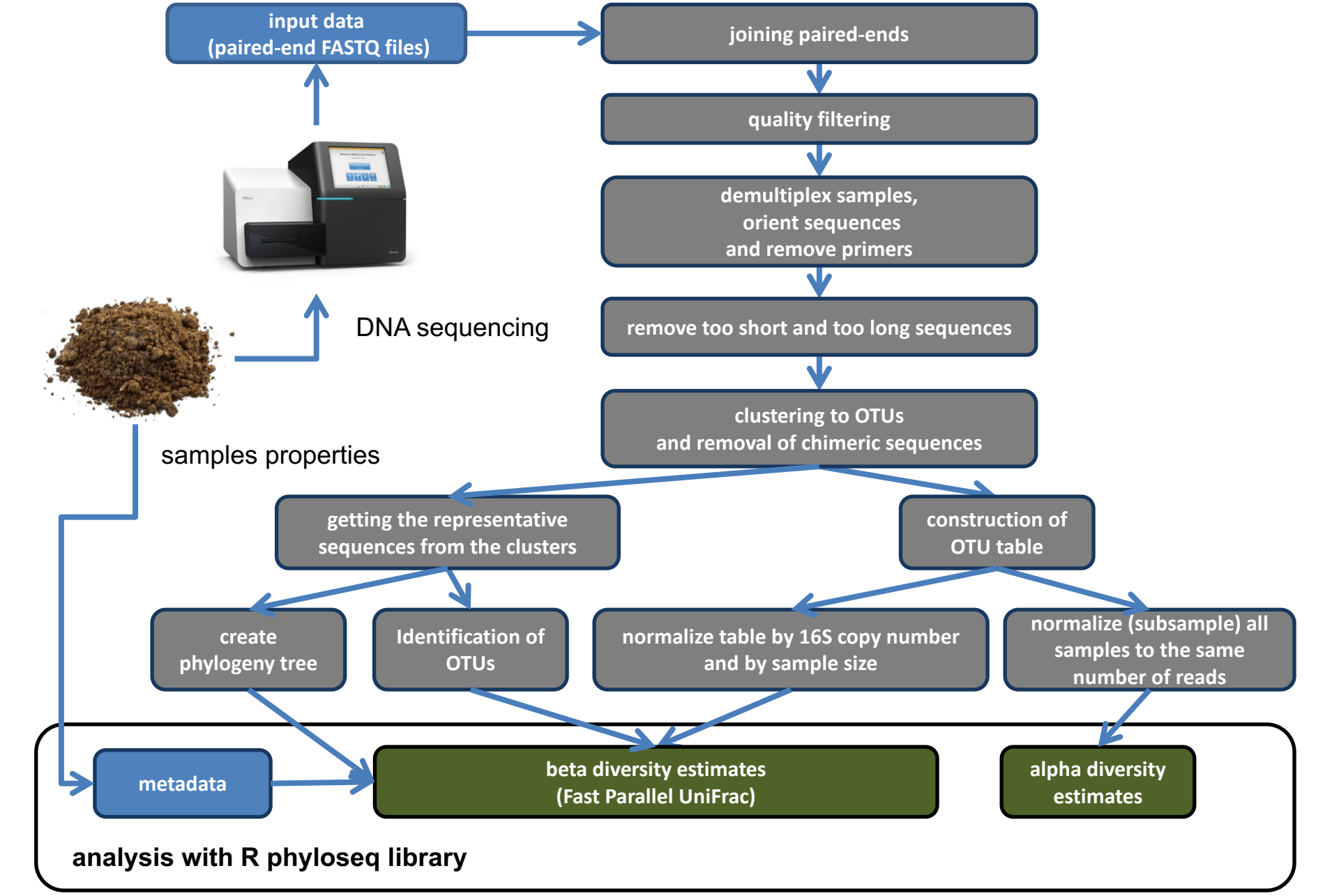
16S rRNA within-genome similarity and copy numbers in bacterial genomes.

Upper panel: the similarity of genomes with various copy numbers: the values indicated represent the first, the second and the third quartile.

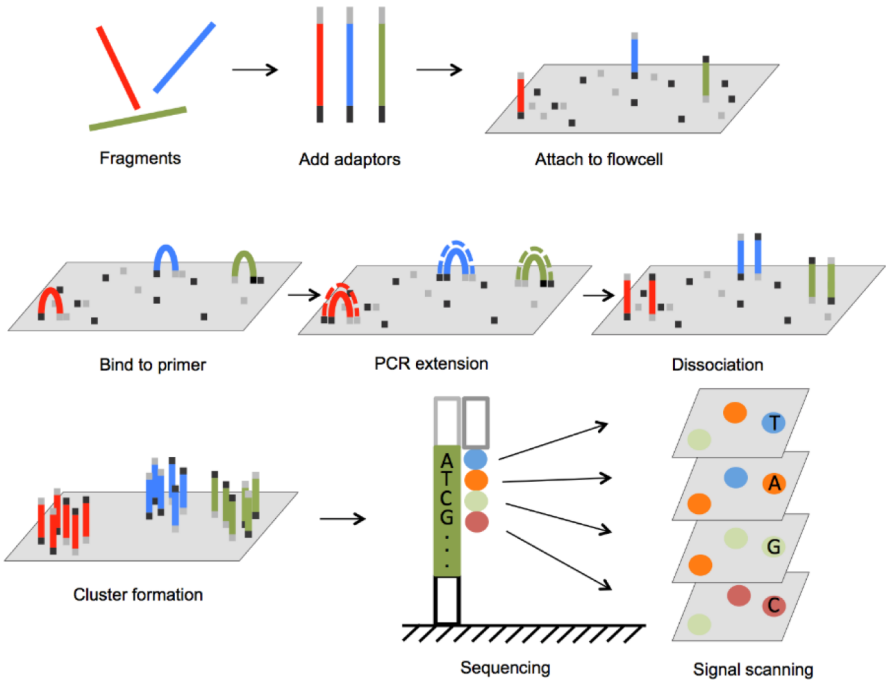
Lower panel: distribution of 16S rRNA copy numbers per genome in 1,690 sequenced bacterial genomes.

T. Větrovský & P. Baldrian - PloS one, 2013

Amplicons pipeline workflow



Input data (paired-end FASTQ files)



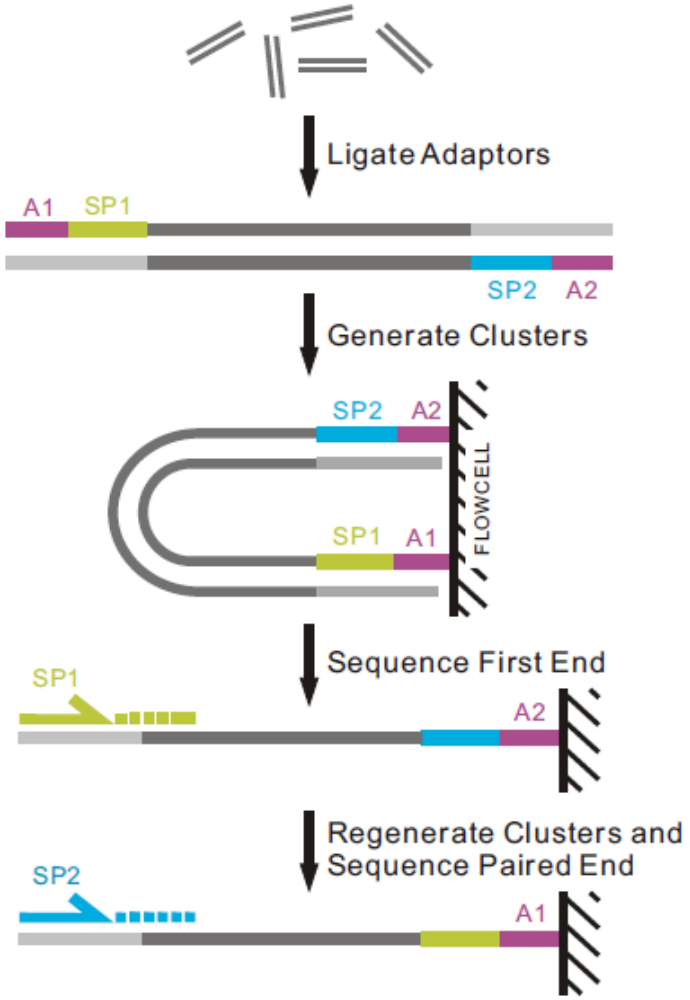
BAC_R1.fastq - first sequence

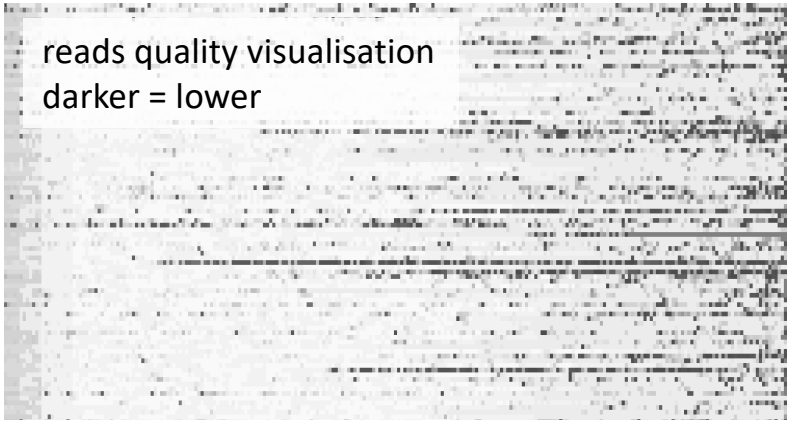
@M03794:8:000000000-AJCUU:1:2114:9990:17907 1:N:0:7
AACAGCCGGACTACTGGGGTTTCTAATCCTGTTTGCTCCCCACGCTTTCGTGC
CTCAGTGTCAATGACCGTGTAGCAAGCTGCCTTCGCAATTGGTGTTCTATGTC
ATATCTAAGCATTTCACCGCTACATGACATATTCCGCTTACCTCCACGATATTC
AAGACTAATAGTATCAATGGCAGTTCCCAAGTTAAGCTCGGGGATTTCACCAC
GGACTTACTAGCCCACCTACGCACCCTTTAAACCCAGT
+
BCCCCFCCCCCCGGGGGGGFGGHHHCHHHHHHHHHHHHHGG2FGGGGGHG
HGGHHHFHHHHHHHHHHHHGGHHHHHHHHHHHHHHHHHGGGGHHHHGHGH
HHHHHHHHFHHHHHHHHGHHHHHHGGGGGHHHGHHHHHHHHHGGGGHHHH
HHHGCGFDHHGGHHHHHHHHGG>GGGHHG/GHGHHHHHHFHFHGHHGHHG?
DGGGGGGGGGGGGACGGGGGGGGGFGGFAAFF;@ADFFFFFFB/FFFFF;

BAC_R2.fastq - first sequence

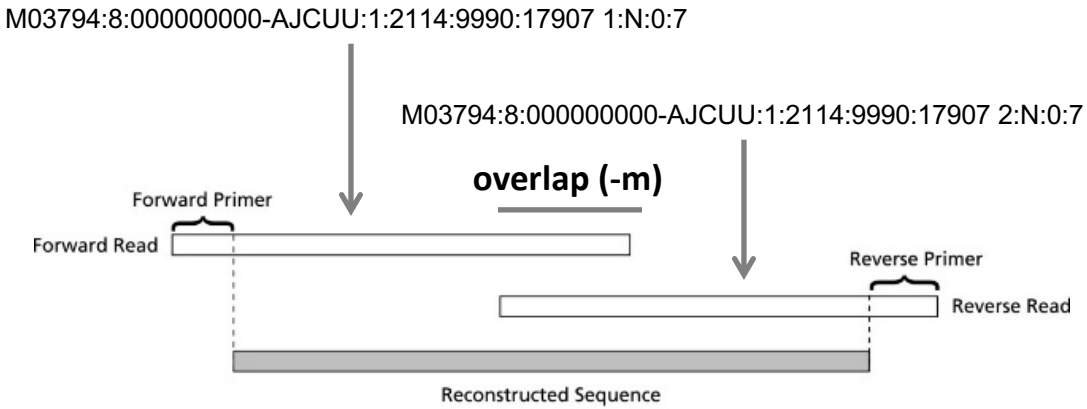
@M03794:8:000000000-AJCUU:1:2114:9990:17907 2:N:0:7
ACGAAGTGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTATCCG
GATTCACTGGGTTTAAAGGGTGCGTAGGTGGGCTAGTAAGTCCGTGGTGAA
ATCCCCGAGCTTAACTTGGGAACTGCCATTGATACTATTAGTCTTGAATATC
GTGGAGGTAAGCGGAATATGTCATGTAGCGGTGAAATGCTTAGATATGACA
TAGAACACCAATTGCGAAGGCAGCTTGCTACACGGTCATTGACACTG
+
BBBBBBBFFBFDGFFFGGGCGGCGGGHHHGDEDGGGFGGHHHGGGGGFDE
E?EEGBGFHHHHGFFGGFFGGFEFGDFDGFFEGGHHHGFEGFHFHEEFFFHH
HHFHGGGFGFHHHFHHHHHHBGHHFBCHFHBGGGHHHHGHHHHHHGFFHH
HGHGD<GGAGHHHGGG@CFHHFCGHH:CCG?AAAGFEFEFGGGBFFFFFFG
FFBBFGFFBDE?BBBB.@9-9..A.B/:AFFFEF.@;AAF//99FFF/

Joining of pair-end data (fastq-join)





reads quality is dropping at the ends




```
fastq-join -v " " -p 15 -m 40 BAC_R1.fastq  BAC_R2.fastq -o bac_joined
```

Quality filtering

$$Q_{illumina} = -10 \times \log_{10}\left(\frac{P_e}{1-P_e}\right),$$

where P_e is the probability of identifying a base incorrectly. For Sanger and other platforms, the formula is as follows [8]:

$$Q_{PHRED} = -10 \times \log_{10}(P_e).$$

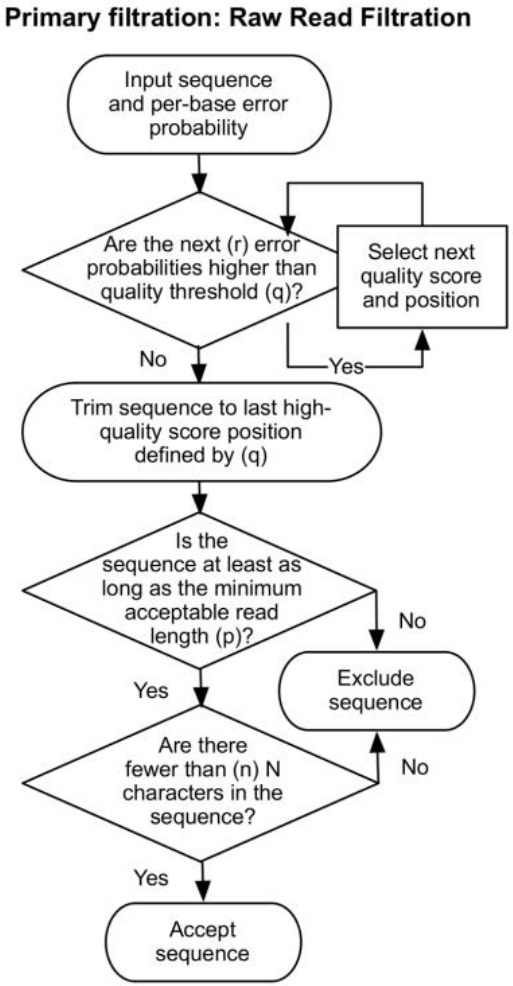
base quality score for Illumina range from 0 to 40

$$Q_{illumina} = 10 \times \log_{10}\left(10^{\left\{\frac{Q_{PHRED}}{10}\right\}} + 1\right)$$

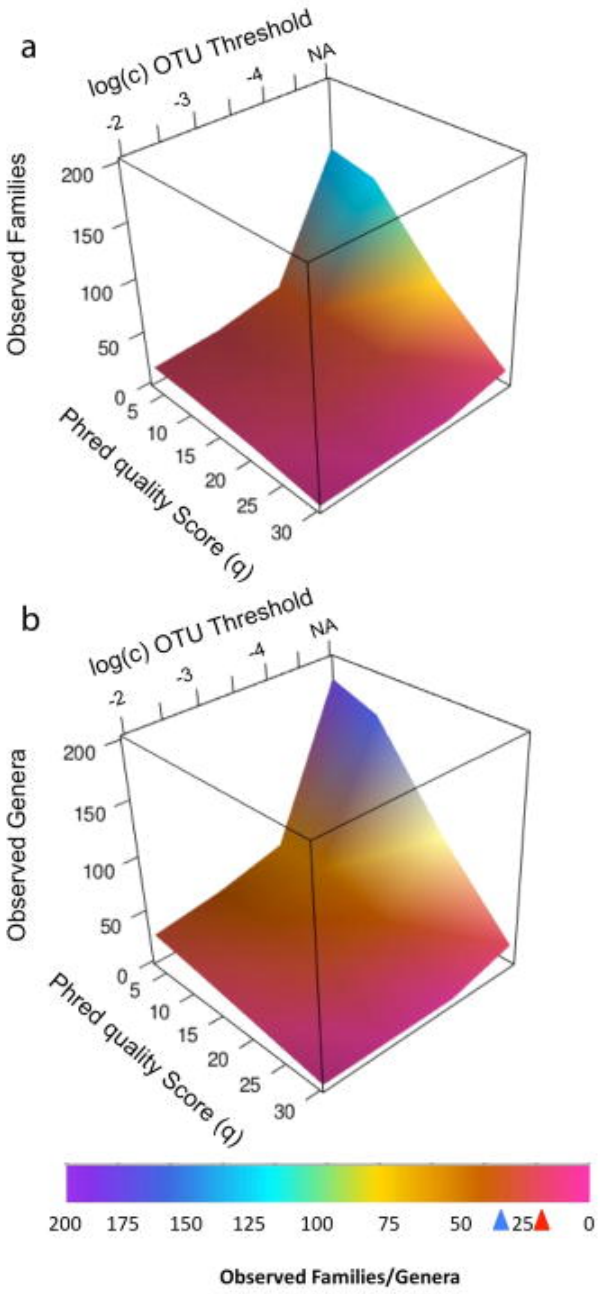
Table 2. Phred quality scores are logarithmically linked to error probabilities (http://en.wikipedia.org/wiki/Phred_quality_score)

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.90%
40	1 in 10 000	99.99%
50	1 in 100 000	99.999%
60	1 in 1 000 000	99.9999%

Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing



α- and β-Diversity comparisons of mock community reads filtered using select phred_quality_score (q) settings (dataset 1). A, B: Family-level (A) and genus-level (B) taxon counts for mock communities filtered with variable (q) values at multiple OTU minimum abundance thresholds (c) (as %).



Quality Filtration Process Flow in QIIME v1.5.0.
Bokulich at al. Nat Methods. 2013 Jan; 10(1): 57–59.

Quality filtering (fastx toolkit)

```
fastq_quality_filter -i bac_joinedjoin -Q33 -q 35 -p 50 -o bac_joinedjoin.qc.fq
```

[-Q33] = Illumina encoded quality scores.
[-q N] = Minimum quality score to keep.
[-p N] = Minimum percent of bases that must have [-q] quality.
[-i INFILE] = FASTA/Q input file. default is STDIN.
[-o OUTFILE] = FASTA/Q output file. default is STDOUT.

FASTQ to FASTA

```
fastq_to_fasta -i bac_joinedjoin.qc.fq -o bac_joinedjoin.qc.fa
```

FASTQ

```
@M03794:8:000000000-AJCUU:1:2114:9990:17907
AACAGCCGGACTACTGGGGTTTCTAATCCTGTTTGCTCCCCAC
GCTTTCGTGCCTCAGTGTCA
+
BCCCCFCCCCCCGGGGGGGFGGHHHCHHHHHHHHHHHHHGG
2FGGGGGHGHGGHHHFHHHHHH
```

FASTA

```
>M03794:8:000000000-AJCUU:1:2114:9990:17907
AACAGCCGGACTACTGGGGTTTCTAATCCTGTTTGCTCCCCACG
CTTTCGTGCCTCAGTGTCA
```

Demultiplexing

each sample is coded by two barcodes

name	FWDprimer	REVprimer
SAMPLE001	515F_T103	806R_T007
SAMPLE002	515F_T002	806R_T052
...		

TAG SPACER ORIGINAL PRIMER

515F_T103 **AATATACGT**GTGCCAGCMGCCGCGGTAA
515F_T002 **ACGAAGT**GTGCCAGCMGCCGCGGTAA

806R_T007 **AGCCACC**GGACTACHVGGGTWTCTAAT
806R_T052 **ATCCTCCC**GGACTACHVGGGTWTCTAAT

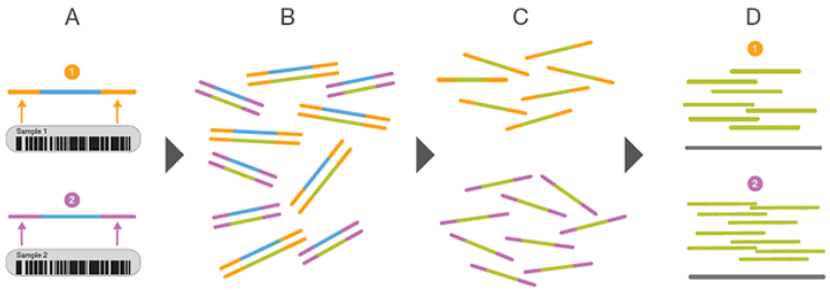
spacer is not presented in native sequences, it is used to prevent overestimation of any taxa

http://press.igsb.anl.gov/earthmicrobiome/protocols-and-standards/16s/

demultiplex_joined_amplicons.py

```
python demultiplex_joined_amplicons.py bac_joinedjoin.qc.fa samples.txt
barcoded_primers_fwd.txt barcoded_primers_rev.txt  bac_joinedjoin.qc.demulti.fa
```

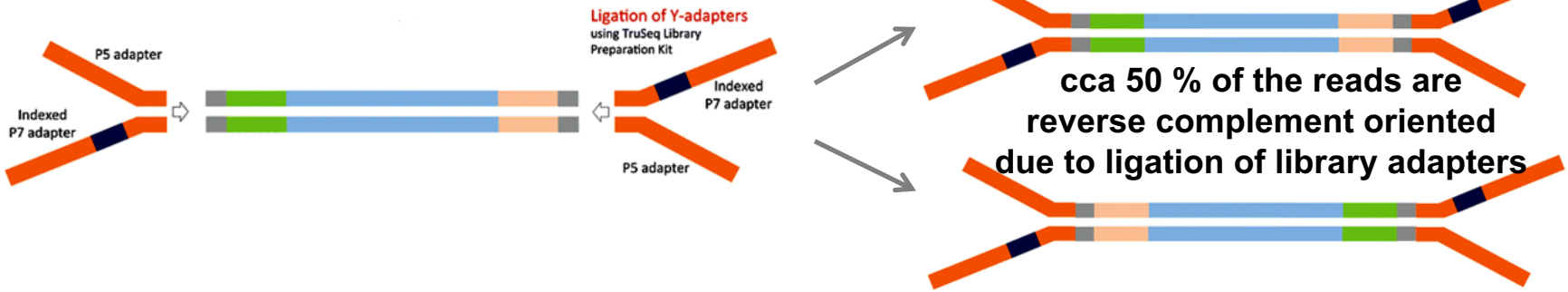
Demultiplex samples



put sample names to sequence titles

>**SAMPLE034**|M03794:8:000000000-AJCUU:1:2114:9990:17907
CCTGTTTGCTCCCCACGCTTTCGTGCCTCAGTGTCAATGACC
GTGTAGCAAGCTGCCTTCGCAATTGGTGTTCTATGTCATATCTA
AGCATTTCACCGCTACATGACATATTCCGCTTACCTCCACGATA
TTCAAGACTAATAGTATCAATGGCAGTTCCCA…

Orient sequences



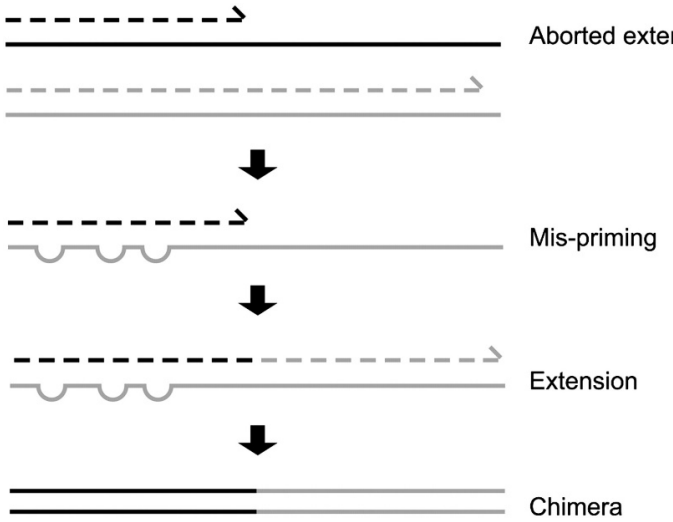
Remove primers

since primer sequences are not native to the sample, they need to be removed before clustering to OTUs

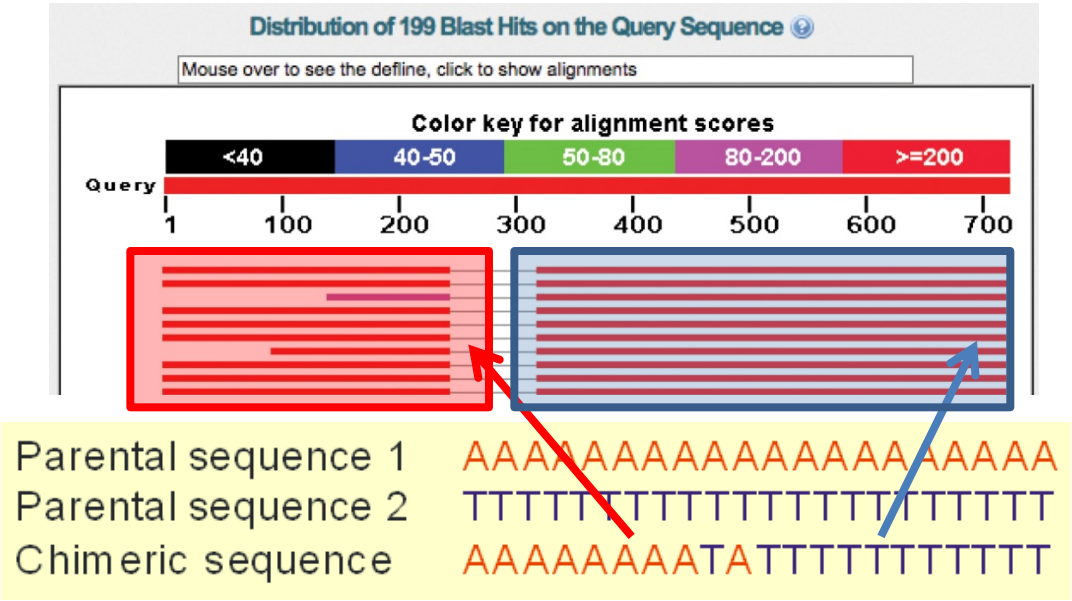
Removing too short and too long sequences (Biopieces)

```
read_fasta -i bac_joinedjoin.qc.demulti.fa | grab -e 'SEQ_LEN >= 245' | grab
-e 'SEQ_LEN <= 260' | write_fasta -x -o bac_joinedjoin.qc.demulti.cut.fa
```

Chimera removal



Formation of chimeric sequences during PCR. An aborted extension product from an earlier cycle of PCR can function as a primer in a subsequent PCR cycle. If this aborted extension product anneals to and primes DNA synthesis from an improper template, a chimeric molecule is formed.



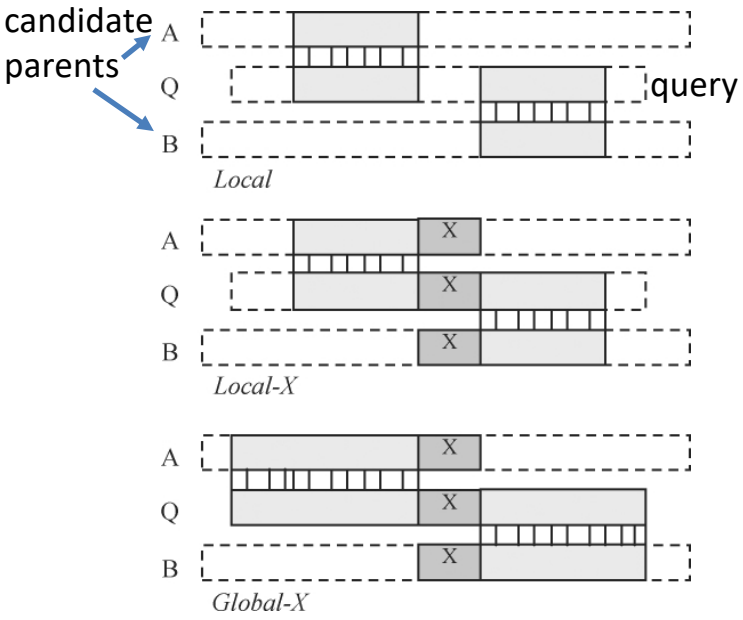
Haas et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011 Mar; 21(3): 494–504.

Chimera removal (UCHIME or realized during clustering to OTUs by UPARSE)

chimeric sequences are derived from parental sequences -> looking for the parents

***de novo* approach:**

- no reference database
- usually based on measuring of several sequence parts abundances – parents should be more abundant then its offspring (chimeras)

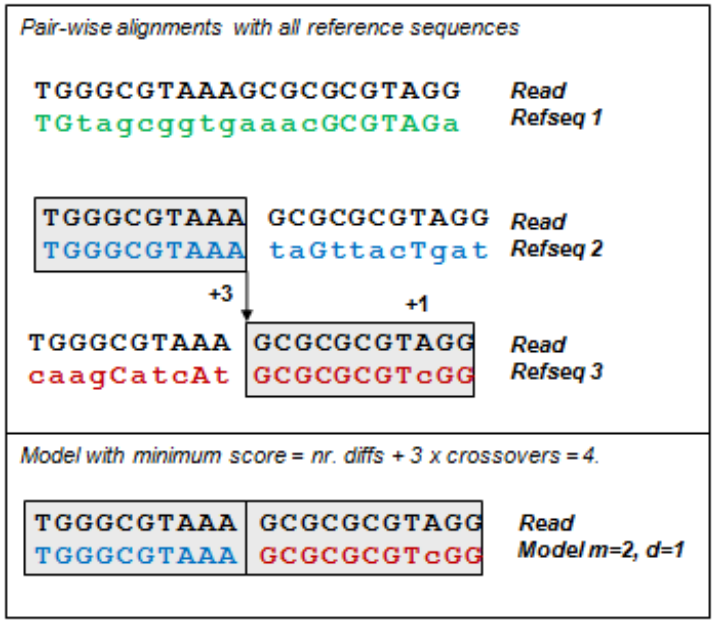


problems

- algorithms are not optimal
- computation cost could be high
- problems with highly similar sequences

reference based approach:

- pair-wise alignment with reference sequence database



problems

- no appropriate reference database for environmental samples
- variable quality of reference databases

Edgar,RC et al. (2011) UCHIME improves sensitivity and speed of chimera detection, *Bioinformatics*

Clustering to OTUs

OTU (Operational taxonomic unit) group of similar sequences grouped based on some similarity threshold usually 97% similarity (16S, ITS) represents Species

Heuristic
- comparison of each sequence with representative sequence („seed“)
- depends on sequence order

FAST

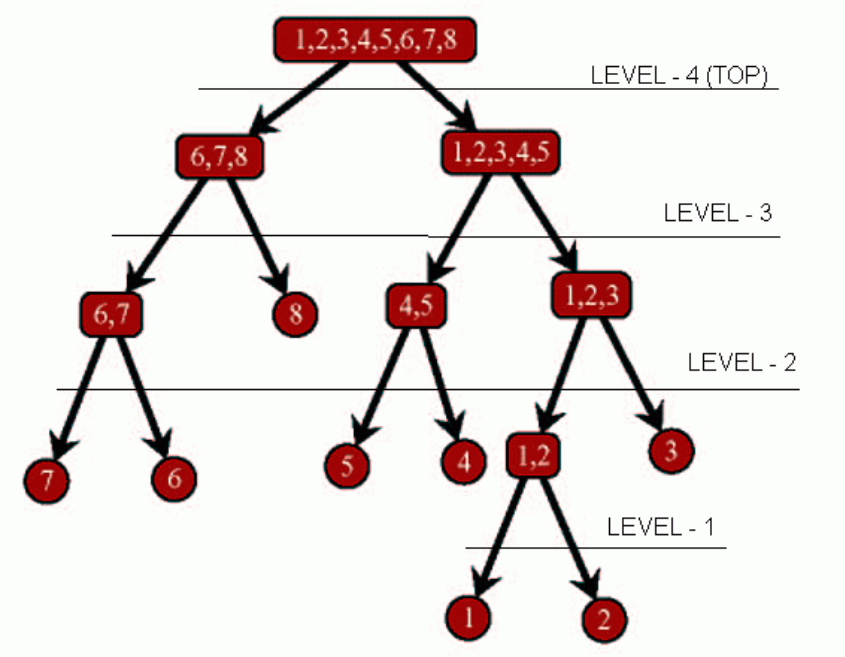
Hierarchical
- comparison of each sequence with each other (tree contruction)

SLOW

Model based
- probabilistic, iterative
- uses more information than the sequence identity

VERY SLOW

Clustering to OTUs (hierarchical)



all pairwise comparisons are performed and OTUs are delineated at fixed distance level

linking method is an important driver of the outcome:

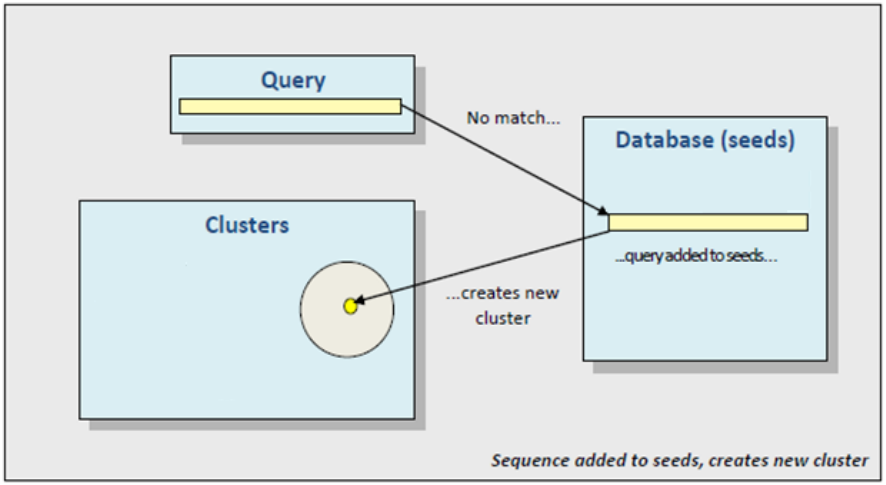
- **single-linkage clustering** (SL) clusters may be merged together due to single sequences being close to each other, even though many of the sequences in each cluster may be very distant to each other
- **complete-linkage clustering** (CL) tends to find compact clusters of approximately equal diameters. With CL, all objects in a cluster are similar to each other
- **average-linkage clustering** (AL) can be seen as an intermediate between single and complete linkage clustering, resulting in more homogeneous clusters than those obtained by the single-linkage method

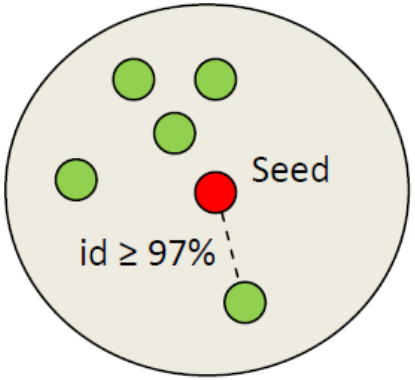
http://iv.slis.indiana.edu/sw/ward.html

Clustering to OTUs (heuristic)

USEARCH

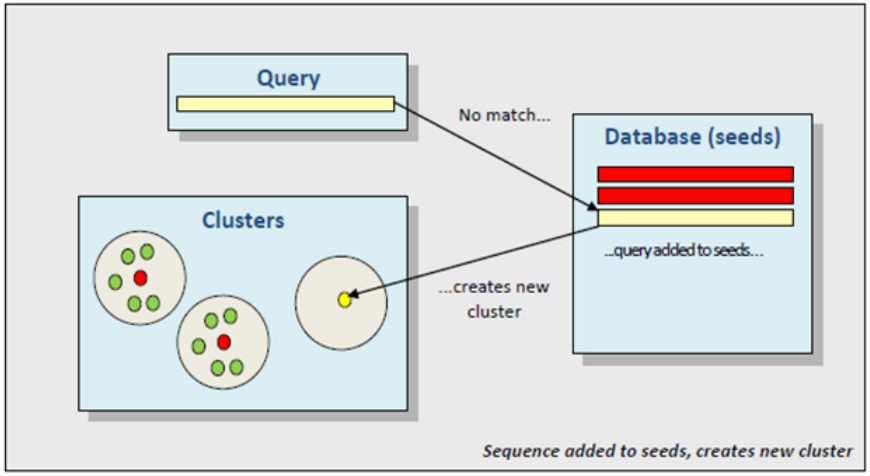
Adding first sequence…



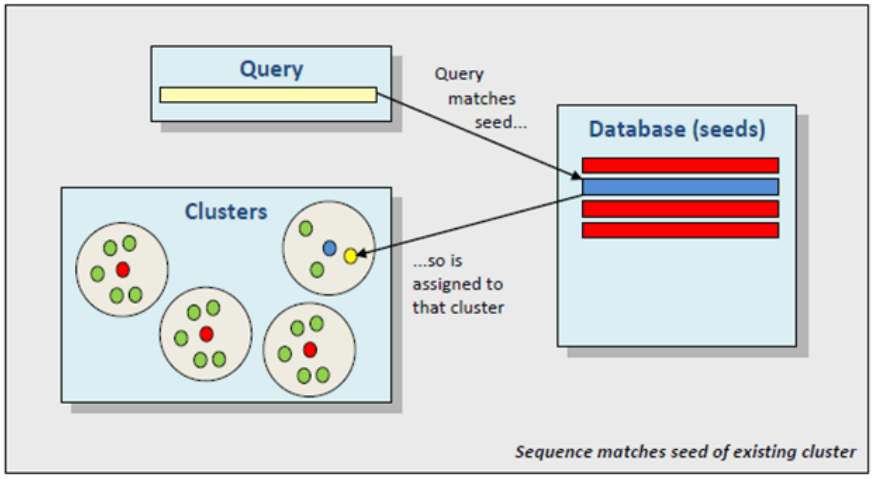


cluster definition

Adding dissimilar sequence…

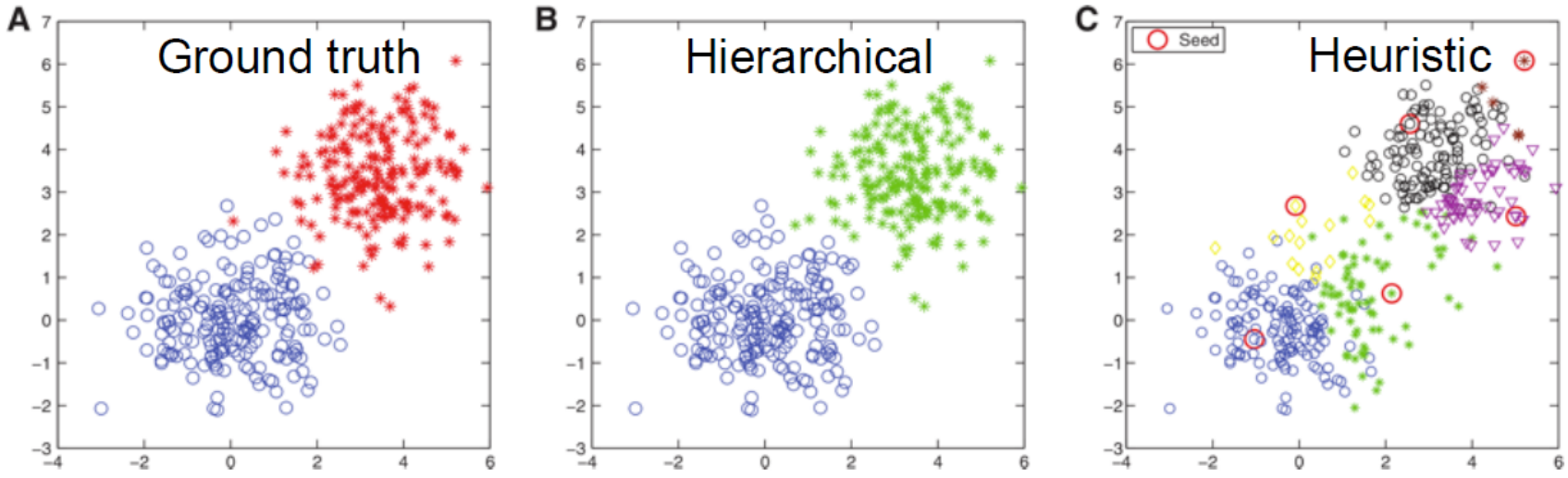


Adding similar sequence…



Edgar,RC (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics* 26(19), 2460-2461.

Clustering to OTUs (hierarchical vs. heuristic)



Hierarchical clustering

- is able to identify the real clusters (ideally)
- computationally expensive

x

Heuristic clustering

- computationally cheap
- often generates artificial clusters (overestimated diversity)

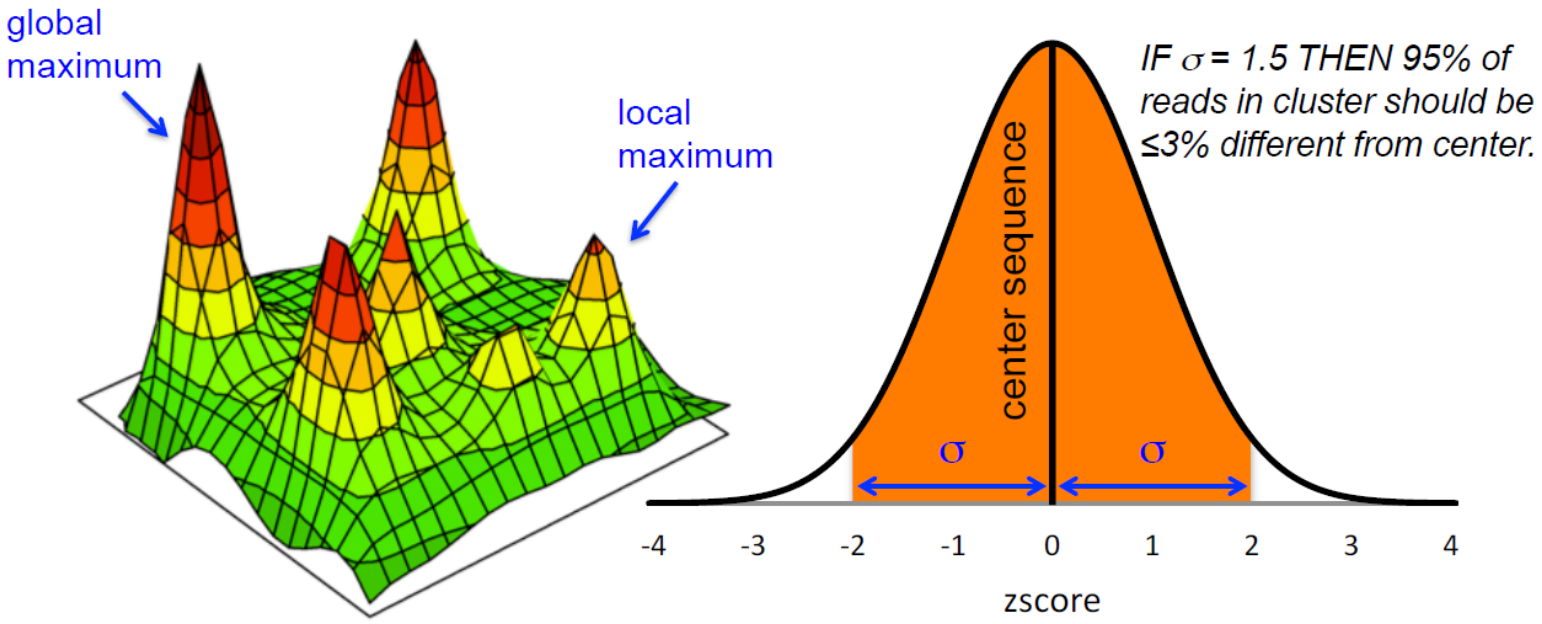
Sun et al. Brief Bioinform. 2012 Jan; 13(1): 107–121.

Clustering to OTUs (model based)

CROP

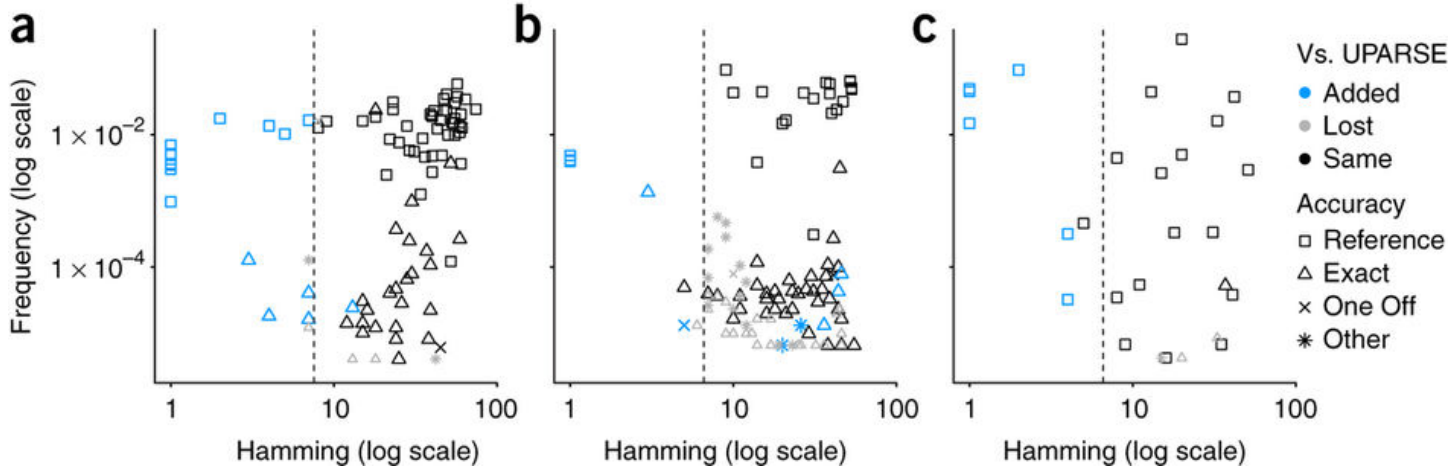
Hao et al. 2011 (Bioinformatics): "If we consider the sequences as data points in a high-dimensional space […], then the probability that a sequence belongs to a cluster becomes a function of the distance between the sequence and the center."

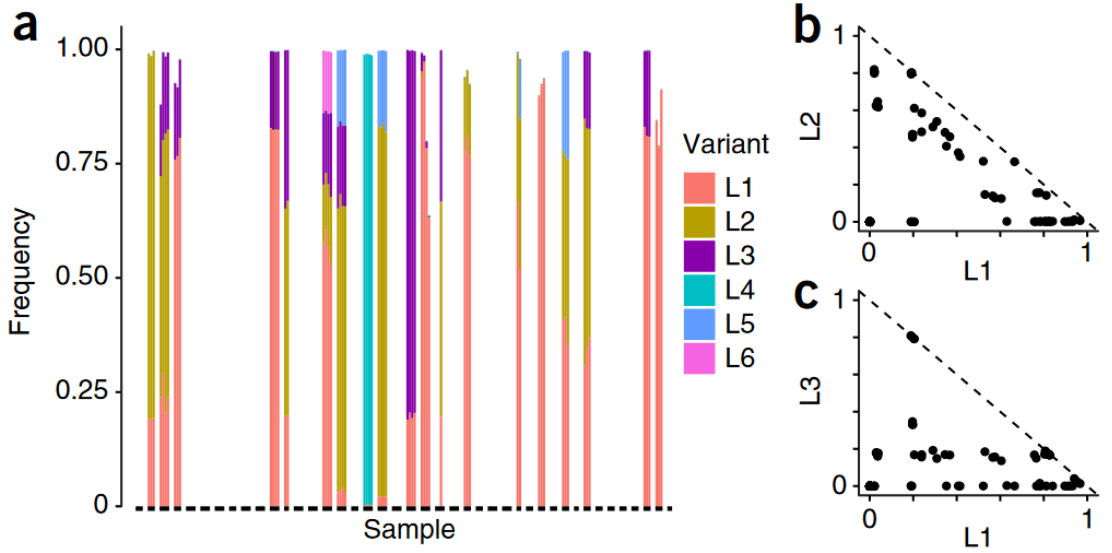
CROP uses a mixture model to find subpopulations among all sequences under the assumption that they are independently drawn from a mixture of Gaussian distributions.



Clustering-independent methods (DADA2)

Comparison of sequence variants inferred by DADA2 with OTUs constructed by UPARSE.





L. crispatus sequence variants in the human vaginal community during pregnancy.
DADA2 identified six L. crispatus 16S rRNA sequence variants present in multiple samples and a significant fraction of all reads.

(a) The frequency of L1–L6 in each sample. Black bars at the bottom link samples from the same subject.
(b,c) The frequency of (b) L1 vs. L2 and (c) L3 vs. L1, by sample. The dashed line indicates a total frequency of 1.

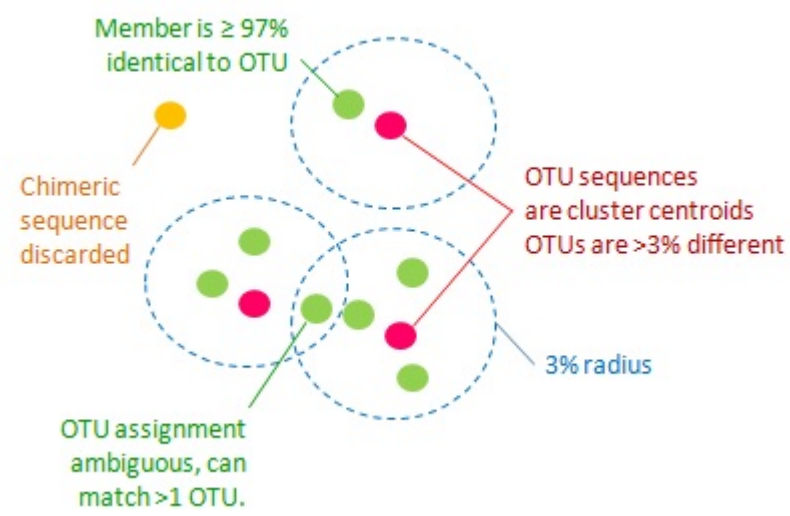
Callahan et al. Nature Methods volume 13, pages 581–583 (2016)

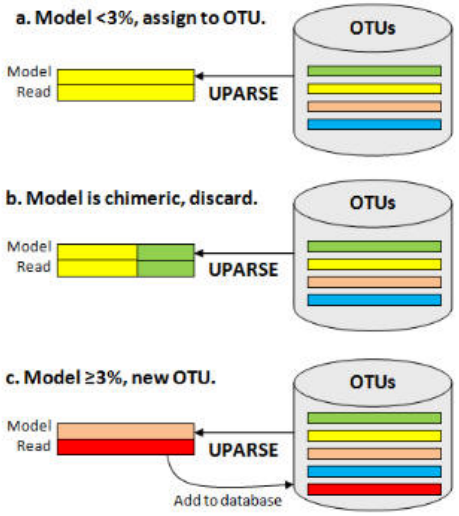
Clustering to OTUs (USEARCH - UPARSE)

```
usearch -fastx_uniques bac_joinedjoin.qc.demulti.cut.fa -fastaout
bac_joinedjoin.qc.demulti.cut.uniques.fa -uc bac_joinedjoin.qc.demulti.cut.uc
-sizeout -relabel Uniq
```

Find the set of unique sequences in an input file, also called dereplication.

```
usearch -cluster_otus bac_joinedjoin.qc.demulti.cut.uniques.fa -otus otus.fa
-relabel Otu
```



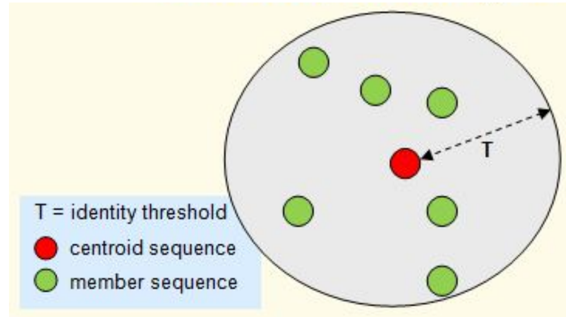


Edgar, R.C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, *Nature Methods*

Getting the representative sequences from the OTUs clusters

centroids of the clusters (seeds)

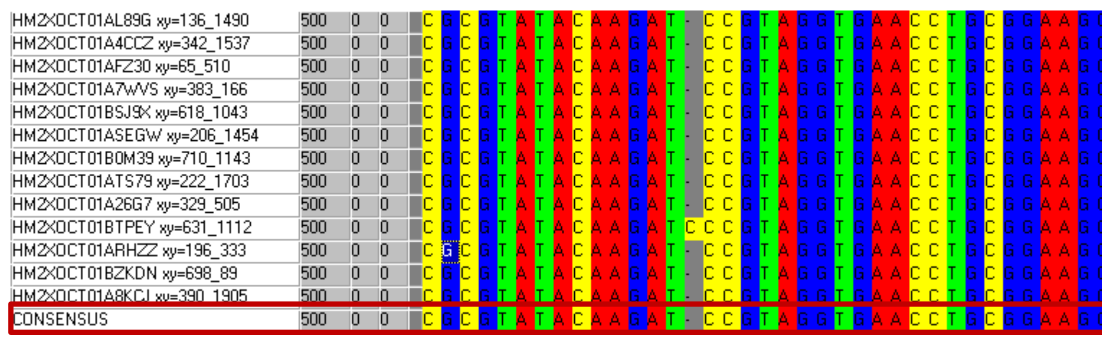
- choosen by clustering algorithm
- might be an outgroup in some cases



otus.fa

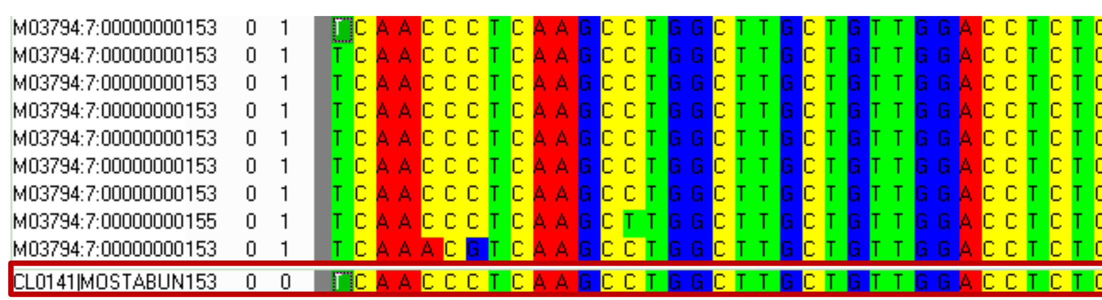
consensual sequence from alignment

- needs multiple alignment of cluster sequences
- good for 454 data



most abundant sequence

- good for Illumina data



Creating an OTU table (USEARCH)

OTU table - matrix that gives the number of reads per sample per OTU

```
usearch -otutab bac_joinedjoin.qc.demulti.cut.fa -otus otus.fa
-otutabout otutab.txt -mapout map.txt
```

#OTU ID	SAMPLE034	SAMPLE025	SAMPLE020	SAMPLE032	SAMPLE009	SAMPLE004	SAMPLE010	SAMPLE005	SAMPLE035	SAMPLE036	SAMPLE017
Otu2	22	8	147	694	472	144	908	180	69	293	28
Otu8	1	0	31	40	1	34	6	0	103	11	103
Otu3	5	0	163	159	3	310	27	2	174	14	188

OTU frequency does not correlate with species frequency
This means, for example, that the most abundant OTU usually does not contain the most abundant species – especially because of multi-copy nature of target genes as 16S and ITS

Singleton counts are especially suspect

- many OTU table entries are often singletons (have value 1) for smaller OTUs because the total count is distributed over several samples
- Small counts are more likely to be spurious, especially singletons, either because the OTU itself is spurious (e.g., an undetected chimera), or because of cross-talk

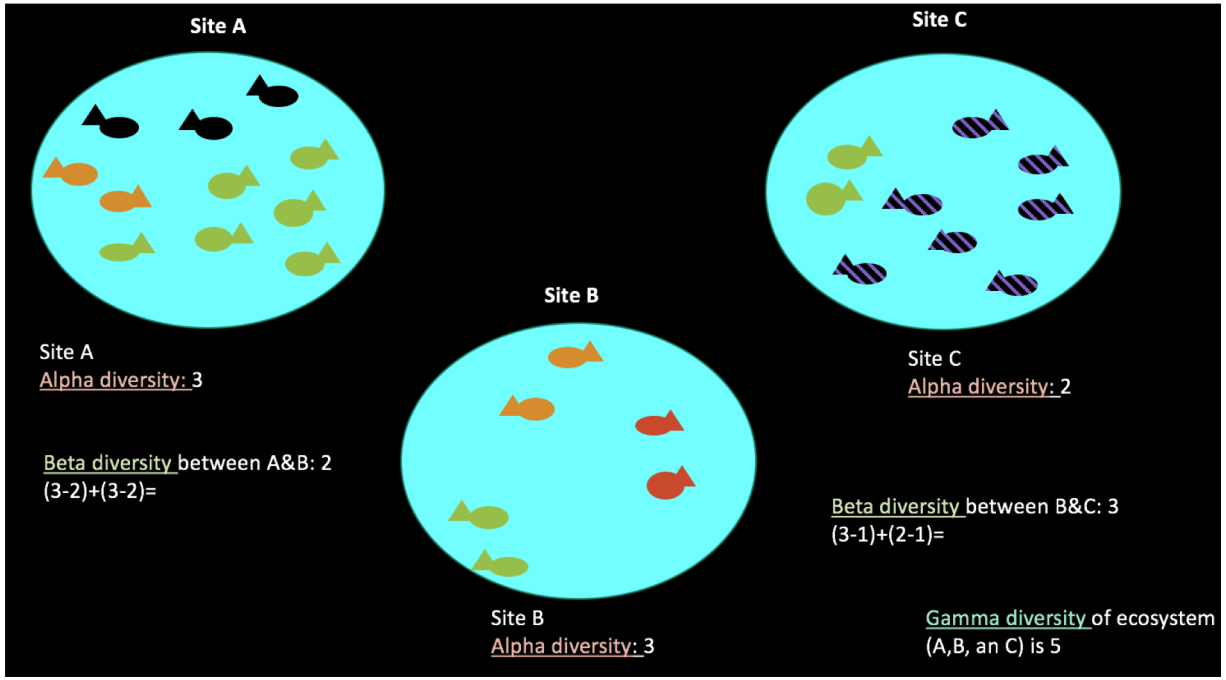
Alpha diversity estimates (UPARSE)

To compare samples diversity all samples need to have same number of reads

```
usearch -otutab_norm otutab.txt -sample_size 5000 -output otutab_norm.txt
```

Alpha diversity is the mean species diversity in sites or habitats at a local scale…
(Diversity of OTUs in individual samples)

```
usearch -alpha_div otutab_norm.txt -output alpha.txt
```



Richness R simply quantifies how many different types the dataset of interest contains. Richness is a simple measure, so it has been a popular diversity index in ecology, where abundance data are often not available for the datasets of interest. Because richness does not take the abundances of the types into account, it is not the same thing as diversity, which does take abundances into account.

Alpha diversity

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

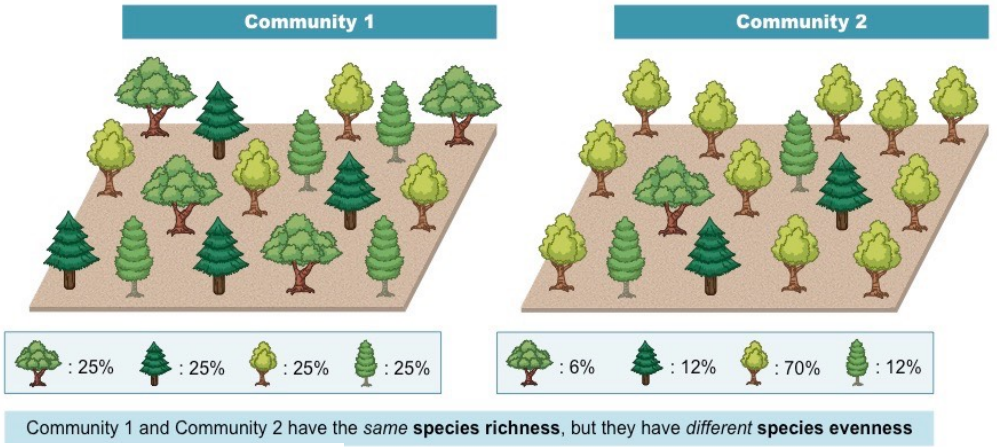
Shannon index (Shannon entropy)

Then the Shannon entropy quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset.

p_i – proportion of the population made up of species i
S – number of species in sample

Species evenness

Species evenness refers to how close in numbers each species in an environment is. Mathematically it is defined as a diversity index, a measure of biodiversity which quantifies how equal the community is numerically.



$$J' = \frac{H'}{H'_{\max}} \qquad H'_{\max} = -\sum_{i=1}^{S} \frac{1}{S} \ln \frac{1}{S} = \ln S$$

Chao1 index

Estimate diversity from abundance data (importance of rare OTUs)

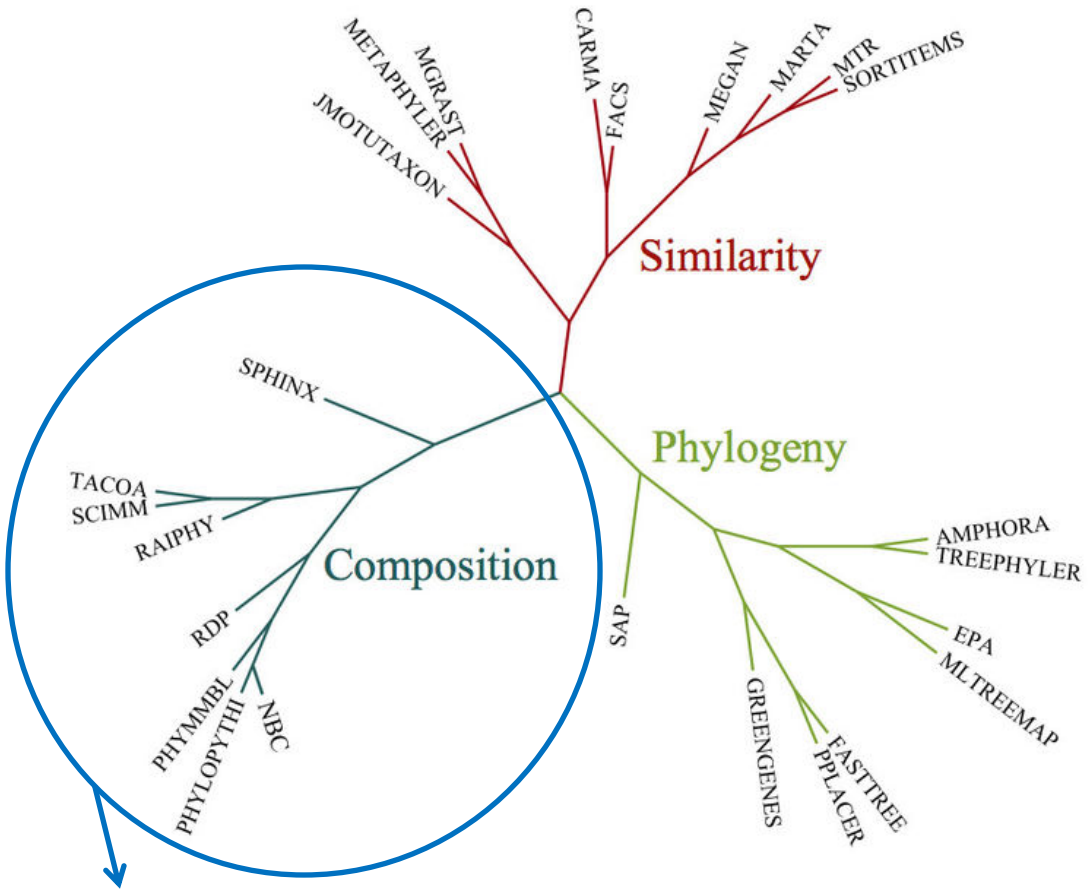
$$S_{est} = S_{obs} + \left(\frac{{f_1}^2}{2f_2}\right)$$

where S_{obs} is the number of species in the sample, f_1 is the number of singletons and f_2 is the number of doubletons.

Shannon, C. E. (1948) A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423 and 623–656.
Chao, A.; Shen, T-J. (2003)

https://palaeo-electronica.org/2011_1/238/estimate.htm

Taxonomic classification of OTUs



Similarity-based

Find homology or minimum alignment distance

Tools:
- local alignments (e.g. BLAST, MEGAN, METAXA2, RTAX)
- global alignments (e.g. GAST)
- overlap alignments (e.g. SINA)

Pro/Con:
- good accuracy for similar sequences
- performs less well on distant lineages
- can be slow on large reference databases

Composition-based

Detect specific features

Tools:
- kmer searches (e.g. NBC/RDP, UTAX, SINTAX)
- hidden Markov models (e.g. PHYMMBL, C16S)

Pro/Con:
- computationally efficient and fast
- performs well on distant lineages
- training required
- limited resolution for shorter sequences

Phylogeny-based

Evolutionary model to determine best placement

Tool:
- ML, NJ, Bayesian (e.g. PPLACER, EPA)

Pro/Con:
- great accuracy for similar sequences
- classification in its evolutionary context
- computationally complex
- requires accurate reference tree
- difficult for non-coding regions

Programs that primarily utilize sequence composition models include Naive Bayes Classifier (NBC)

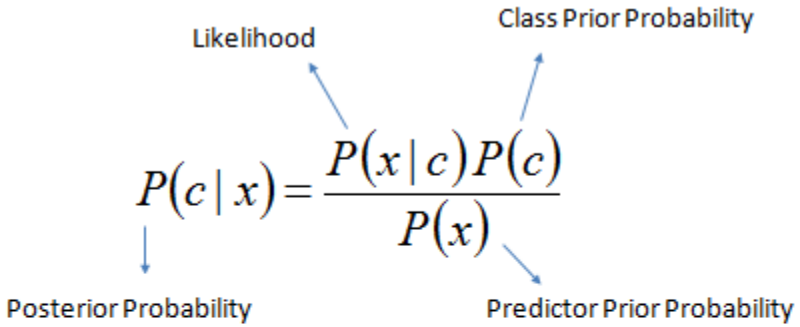
- **To date, no algorithm is convincingly outperforming the others**
- **Other facors suchas genetic markers/locus, sequence quality, and integrity of reference database are likely more crucial**

```
java -Xmx1g -jar ~/RDPTools/RDPTools/classifier.jar classify -g 16srrna -c 0.8
-o otus_classified.txt -h otus_hier.txt otus.fa
```

Bazinet & Cummings, A comparative evaluation of sequence classification programs, BMC Bioinformatics 2012 13:92

Naive Bayes Classifier

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

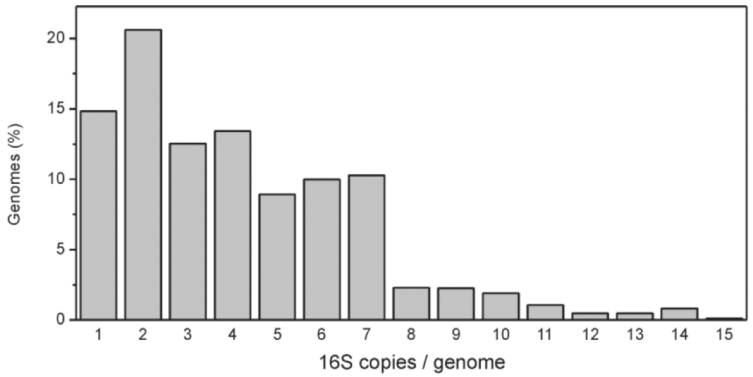
- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

In ZeroR model there is no predictor, in OneR model we try to find the single best predictor, naive Bayesian includes all predictors using Bayes' rule and the independence assumptions between predictors.

http://www.saedsayad.com/naive_bayesian.htm

Normalize OTU table by 16S copy number

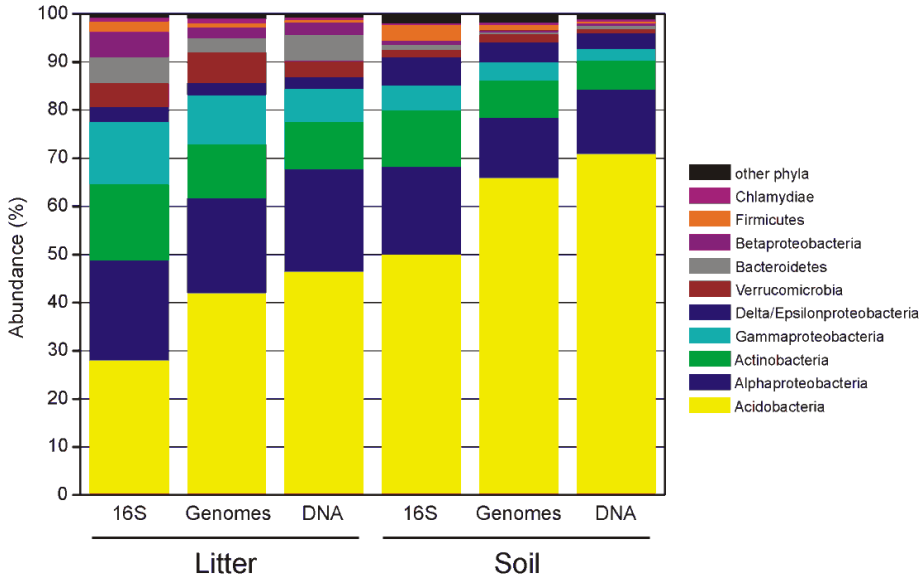
```
python 16S_copy_numbers_and_taxomony.py 0.8 otus_classified.txt rrnDB-5.4_pantaxa_stats_RDP.tsv otus.taxonomy.txt
```



rrnDB

A searchable database documenting variation in ribosomal RNA operons (rrn) in Bacteria and Archaea. Find information such as the 16S gene copy number of an organism by looking up its name under the NCBI or RDP taxonomy or by full-text search of rrnDB's records.

```
python normalise_by_16Scopy_and_samples.py otutab.txt otus.taxonomy.txt otutab.16Snorm.txt
```



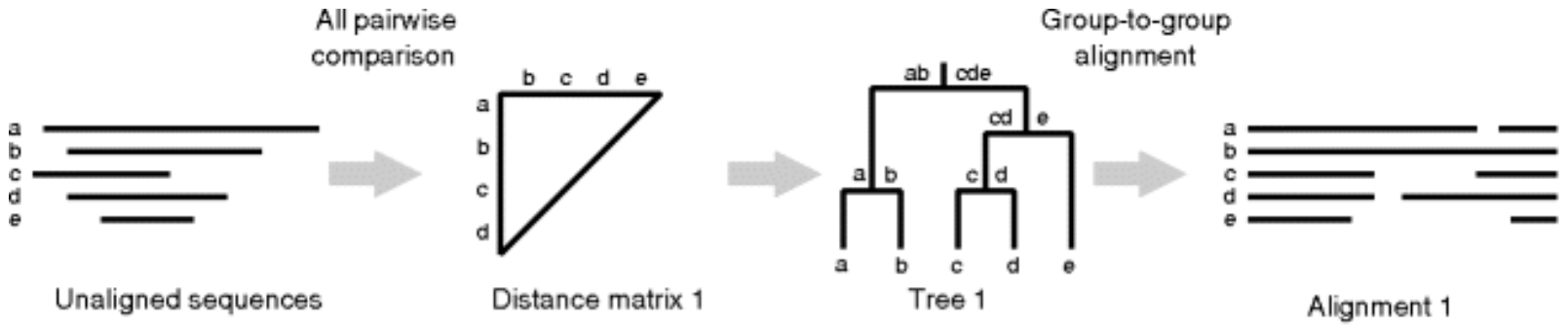
Abundance of bacterial 16S rRNA sequences, genomes and DNA in forest litter and soil.

Relative abundance of bacterial 16S rRNA sequences in the amplicon pool from Picea abies litter and soil (Baldrian et al., 2012), and estimates of the relative abundance of bacterial genomes and DNA. The estimates were calculated using the values of 16S rRNA copy numbers and genome sizes of the closest hits to each bacterial OTU.

T. Větrovský & P. Baldrian - PloS one, 2013 & Stoddard et al. (2015) https://rrndb.umms.med.umich.edu/

create phylogeny tree - multiple alignment & neighbor joining (mafft & fastphylo)

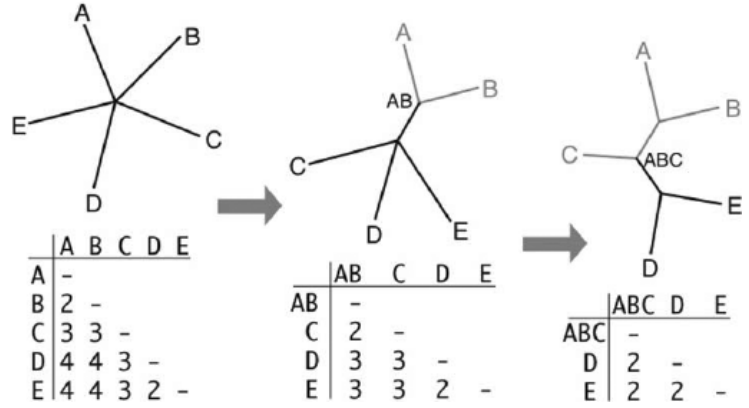
```
mafft --inputorder --auto otus.fa > otus.aligned.fa
```




```
fastdist  -o otus.aligned.xml  -D JC -I fasta otus.aligned.fa
```

computes distance matrices out of multialignments

```
fnj  -o otus_tree.nwk -O newick -m BIONJ -b 0 otus.aligned.xml
```



Neighbor Joining algorithm: start from a star phylogeny (left); find the nearest pair of nodes (according to the distance matrix, either of A-B or D-E) (middle); recalcu- late the distance matrix using the new node (AB); repeat until the tree is fully resolved (right).

Kazutaka Katoh et al. (2009) Bioinformatics for DNA Sequence Analysis pp 39-64
Saitou & Nei (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees.

Beta diversity

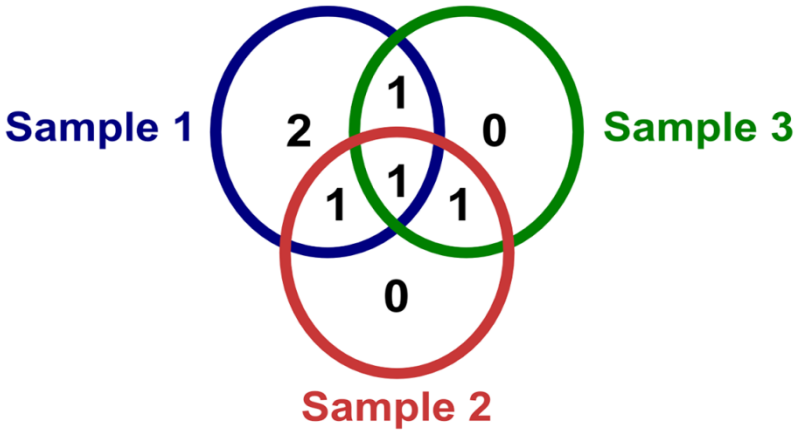
These examples describe the A) sequence counts and B) relative abundances of six taxa (A, B, C, D, E, and F) detected in three samples.

A) Sequence Abundance

OTU	Sample 1	Sample 2	Sample 3
A	60	0	35
B	24	5	5
C	10	0	0
D	5	0	0
E	1	0	0
F	0	20	10
Total	100	25	50

B) Sequence Relative Abundance

OTU	Sample 1	Sample 2	Sample 3
A	0.60	0	0.70
B	0.24	0.20	0.10
C	0.10	0	0
D	0.05	0	0
E	0.01	0	0
F	0	0.80	0.20
Total	1.0	1.0	1.0



Beta diversity represents the similarity (or difference) in organismal composition between samples. In this example, it can be simplistically defined by the equation:

$$\beta = (n_1 - c) + (n_2 - c)$$

where n_1 and n_2 are the number of taxa in samples 1 and 2, respectively, and *c* is the number of shared taxa, but again many metrics such as Bray-Curtis or UniFrac are commonly employed.

Morgan & Huttenhower (2012) Chapter 12: Human Microbiome Analysis. PLOS Comp. Biol. https://doi.org/10.1371/journal.pcbi.1002808

Beta diversity

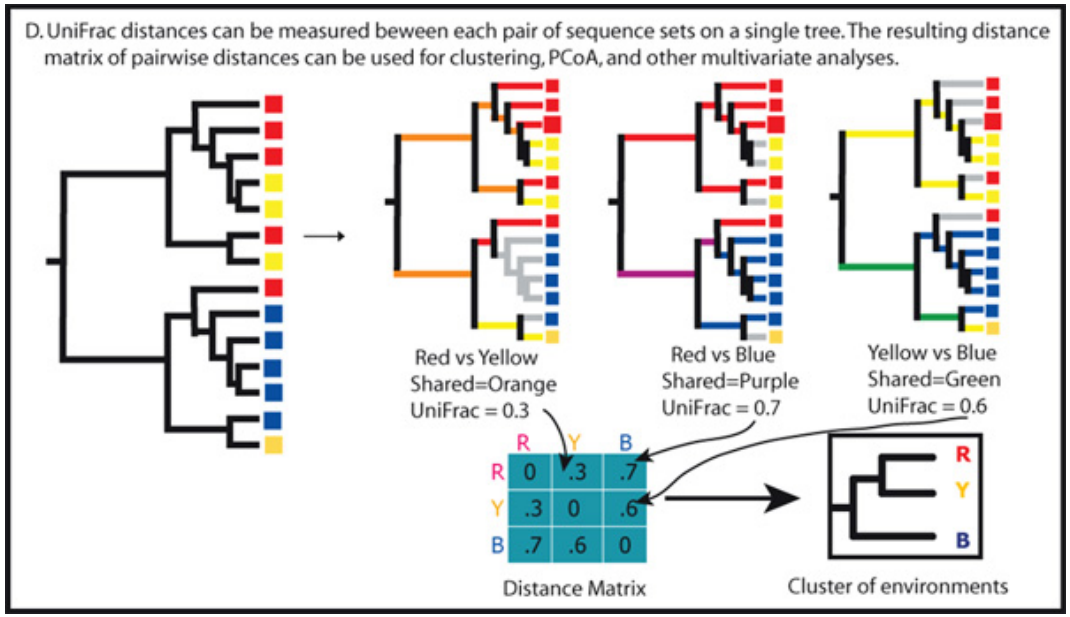
Bray–Curtis dissimilarity
is used to quantify the differences in species populations between two different sites. The Bray-Curtis dissimilarity is always a number between 0 and 1. If 0, the two sites share all the same species; if 1, they don't share any species.

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

i & *j* are the two sites,
S_i is the total number of specimens counted on site i,
S_j is the total number of specimens counted on site j,
C_{ij} is the sum of only the lesser counts for each species found in both sites.

UniFrac

Is a technique used to compare how similar two organisms share their phylogenetic information. In this process the two organisms are labeled either red or blue. The two are then compared each other using either a specifically colored shared branch or a specifically colored unshared branch leading to a specific taxon. The number of unshared branched are then divided by the total number of branches to find the distance between two taxa. The closer the number is to 0, the more similar the two taxa are to each other. The distance is only 0 when the two taxa are identical. This system satifies the distance metric.

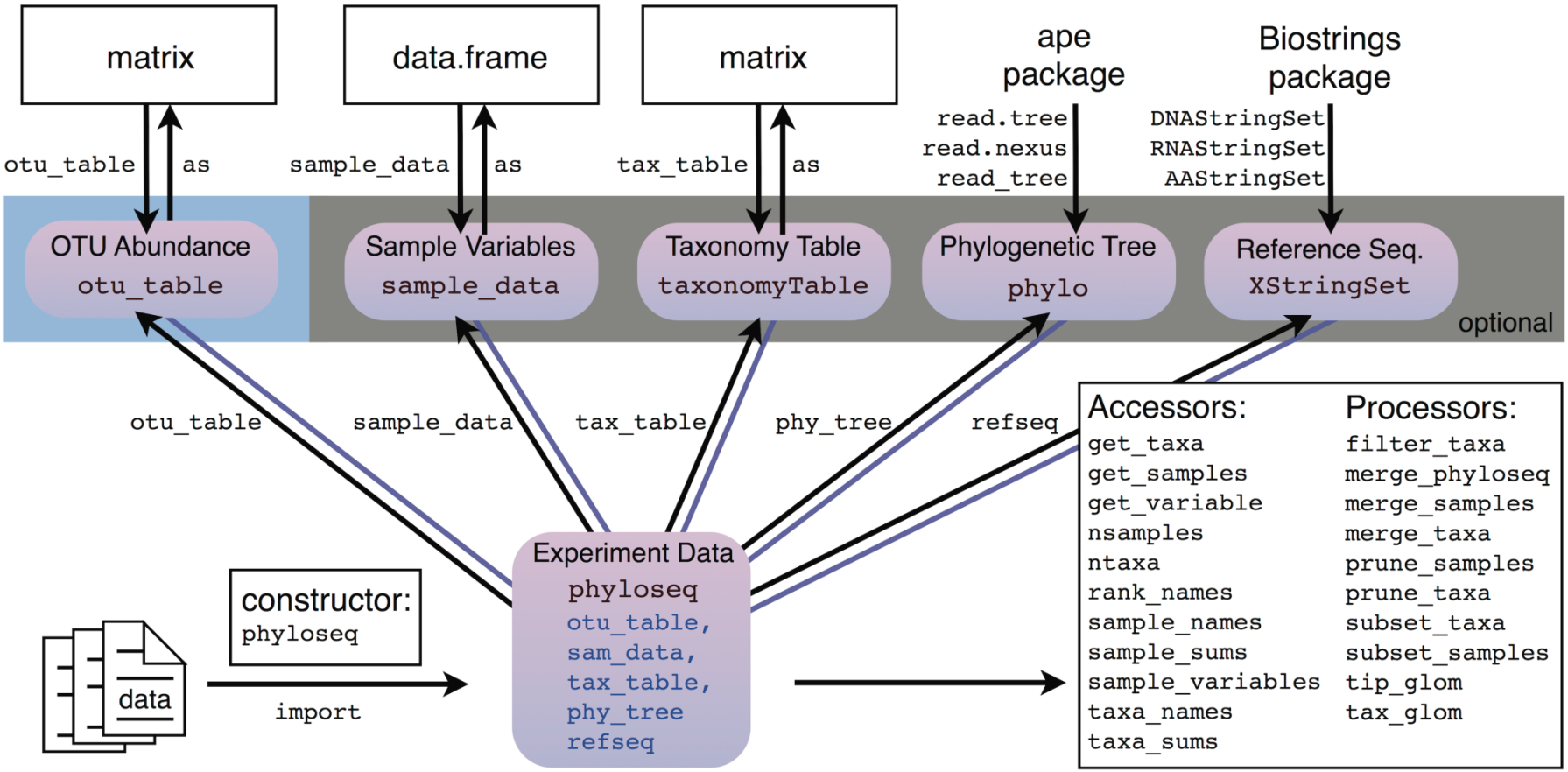


http://www.statisticshowto.com/bray-curtis-dissimilarity/

Lozupone, C.; Knight, R. (2005). "UniFrac: A New Phylogenetic Method for Comparing Microbial Communities". Applied and Environmental Microbiology. 71 (12): 8228–8235.

Analysis using R phyloseq library

The phyloseq package is a tool to import, store, analyze, and graphically display complex phylogenetic sequencing data that has already been clustered into Operational Taxonomic Units (OTUs), especially when there is associated sample data, phylogenetic tree, and/or taxonomic assignment of the OTUs.

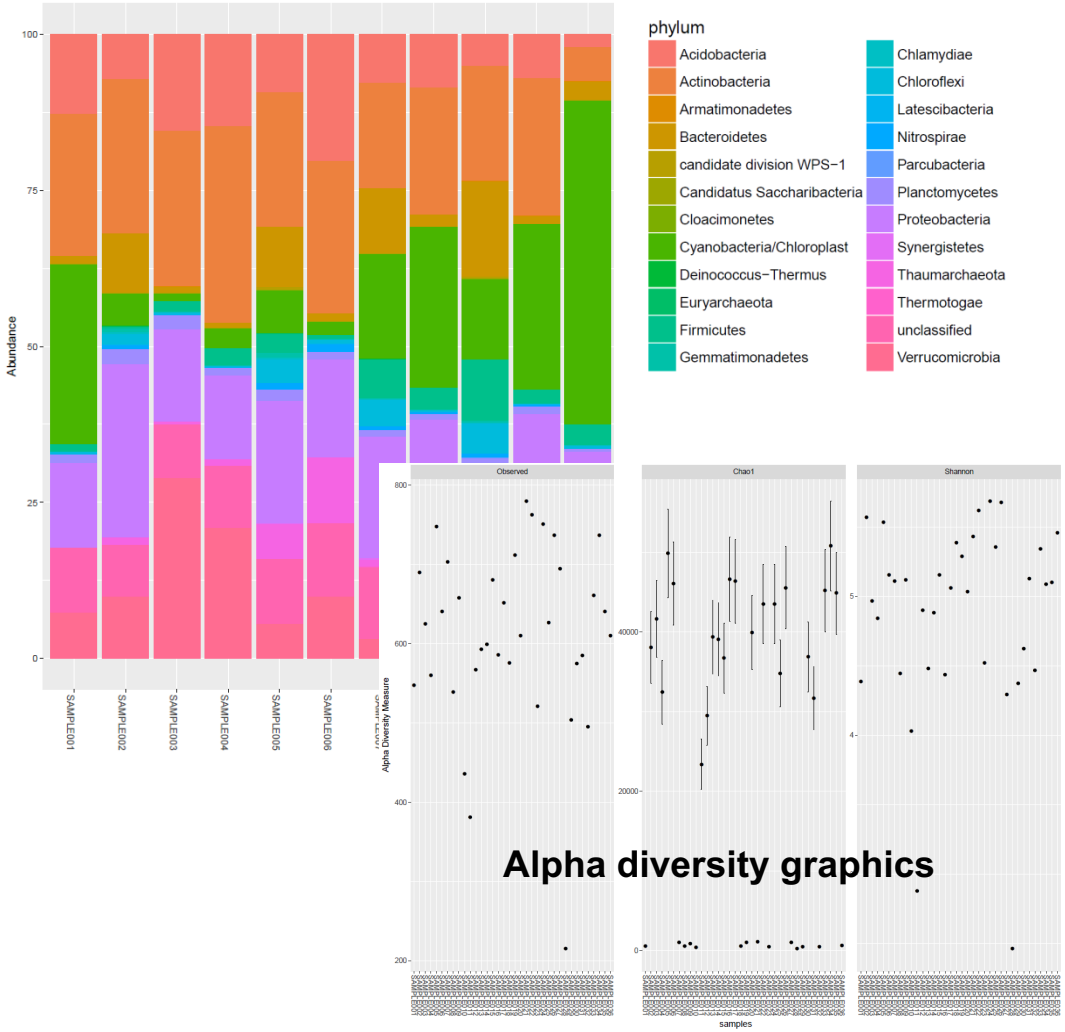


The phyloseq class is an experiment-level data storage class defined by the phyloseq package for representing phylogenetic sequencing data.

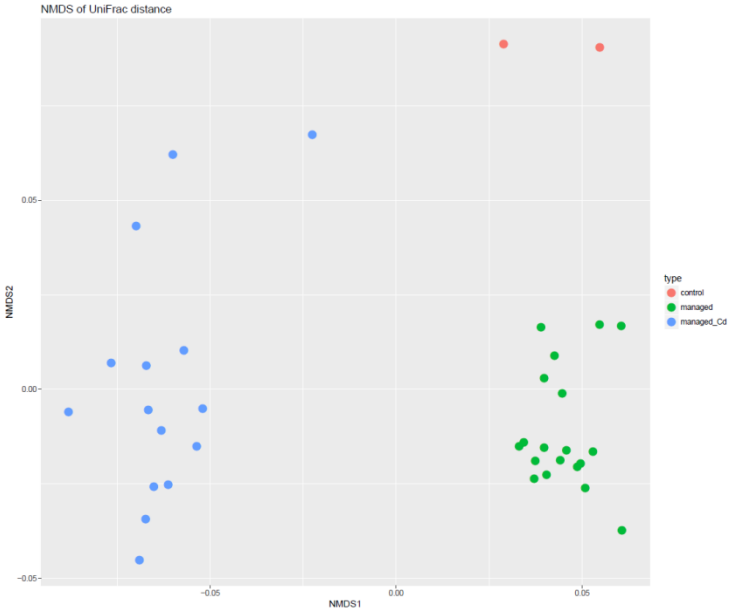
McMurdie and Holmes (2013) phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data PLoS ONE 8(4):e61217

Analysis of the results using R phyloseq library (http://joey711.github.io/phyloseq/)

Phylogeny plot bars



Ordination Plots (Fast Parallel UniFrac)

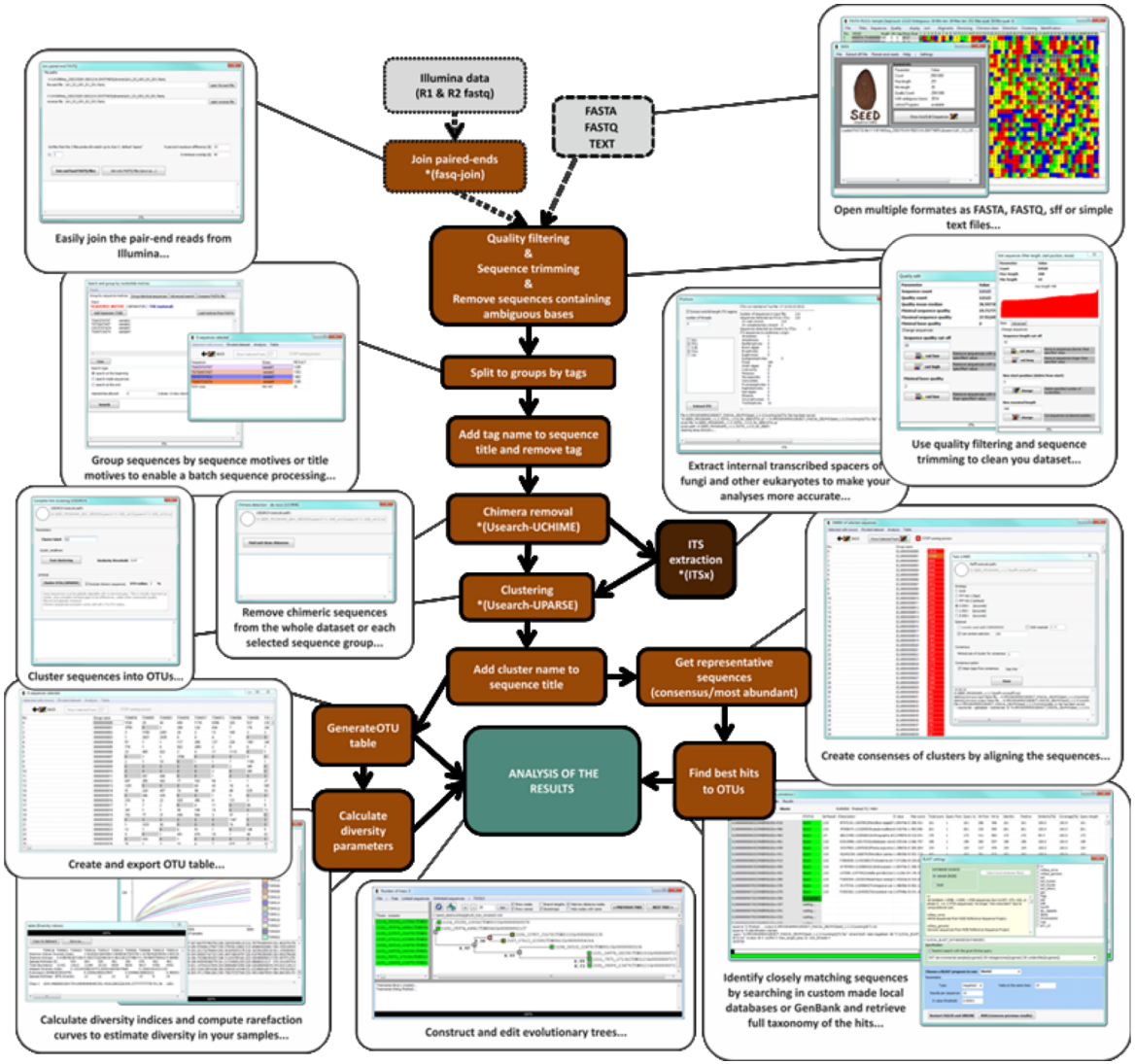


Alpha diversity graphics

…and more:
Plot Microbiome Network
Heatmap Plots
Trees

McMurdie and Holmes (2013) phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data PLoS ONE 8(4):e61217

GUI based alternative for Windows (http://www.biomed.cas.cz/mbu/lbwrf/seed/)



Větrovský, Baldrian & Morais (2018) SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. Bioinformatics, bty071, 2018

SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses

- editing of sequences and their titles
- sorting
- quality trimming
- pair-end joining
- grouping of sequences based on sequence motifs or sequence titles
- batch processing of sequence groups
- denoising
- chimera removal
- ITS extraction
- sequence alignments and clustering
- OTU table construction
- construction of consensus sequences
- creation of local databases for BLAST
- searching either local databases or the whole NCBI
- retrieval of taxonomical classification from the NCBI
- calculation of diversity parameters
- many more...